2013

Investigations into Visual Statistical Inference

Md Mahbubul Amin Majumder
Iowa State University

Investigations into Visual Statistical Inference

by

Md. Mahbubul Amin Majumder

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:

Dianne Cook, Major Professor

Heike Hofmann, Major Professor

Frederick O. Lorenz

Philip Dixon

Leslie Miller

Michelle Graham

Iowa State University

Ames, Iowa

2013

DEDICATION

To my adorable daughter, Azra, who left us just before her first birthday.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank my major professors, Dr. Dianne Cook and Dr. Heike Hofmann, for their valuable guidance in my research work. The lessons I learnt on data visualization from them were very important in conducting this research. They also extended their support during my difficult times.

Dr. Cook provided invaluable counseling and encouragements to overcome my shocks after I had lost my daughter. She has been always very watchful and at times went beyond the academic limit. That helped me reach to a marginal point just to be in track and finally finish this work.

My committee members provided valuable comments and suggestions. I appreciate their times in reviewing my dissertation.

I would like to thank John Hobbs for mentoring me during my academic study at Iowa State University. I also thank the Graphics group of the department of statistics, who voluntarily participated in my pilot studies and provided me with various suggestions.

This work was funded in part by National Science Foundation grant DMS 1007697. The department of statistics provided assistantship as well. I am very thankful to have these opportunities.

Finally I thank my parents and family who constantly provided various support during my academic endeavor.

ABSTRACT

Statistical graphics play an important role in exploratory data analysis, model checking and diagnostics, but they are not usually associated with statistical inference. Recent developments allows inference to be applied to statistical graphics. A new method, called the lineup protocol, enables the data plot to be compared with null plots, in order to obtain estimates of statistical significance of structure. With the lineup protocol observed patterns visible in the data can be formally tested. The research conducted and described in this thesis validates the lineup protocol, examines the effects of human factors in the application of the protocol, and explains how to implement the protocol. It bridges the long existing gulf between exploratory and inferential statistics. In the validation work, additional refinement of the lineup protocol was made: methods for obtaining the power of visual tests, and p-values for particular tests are provided. A head-to-head comparison of visual inference against the best available conventional test is run for regression slope inference, using simulation experiments with human subjects. Results indicate that the visual test power is higher than the conventional test when the effect size is large, and even for smaller effect sizes, there may be some super-visual individuals who yield better performance than a conventional test. The factors that may influence the individual abilities are examined, and results suggest that demographic and geographic factors have statistically significant but practically insignificant impact. This work provides instructions on how to design human subject experiments to use Amazon's Mechanical Turk to implement the lineup protocol.

CHAPTER 1. INTRODUCTION

Visualization of data is an important part of statistical data analysis. It is used for discovering structure in data, initial data analysis and model checking. It is not usually associated with statistical inference. Recent developments in statistical graphics allows it to be used as a tool for statistical inference. With the introduction of lineup protocol the hypothesized pattern in the data can be formally tested using statistical graphics. This thesis work focuses on validating the lineup protocol and further developing visual statistical inference technique defining necessary terminologies. It bridges the long existing gulf between exploratory and inferential statistics.

The dissertation is organized as three independent papers. The first paper in Chapter 2 defines necessary terms to develop visual statistical inference techniques, presents the methods to obtain the power of visual test and compares the power with that of conventional test. The second paper in Chapter 3 presents the human factors that may affect the performance of the lineup protocol and examines their influence on the human observer who evaluates the lineup. The final paper in Chapter 4 describes how to design human subject experiments and develops a web application to get the lineups evaluated by online observers recruited through Amazon Mechanical Turk web site.

1.1 Overview

The concept of lineup protocol is brought from the police lineup where the suspect is aligned together with some innocent people. The null hypothesis is that the suspect is not guilty until proven. If the witness can identify the suspect from the lineup the null hypothesis is rejected. Buja et al. (2009) proposed the similar idea of police lineup in testing the discoveries made from

the exploratory data analysis. A plot of the observed data or the suspect is placed in a layout of plots called lineup where the rest of the plots in the lineup, called null plots, are generated from the model specified by the null hypothesis. An observer is asked to evaluate a lineup to see if he or she can correctly pick the observed plot and based on that the null hypothesis is rejected. Whether this method works or to what extent it works are some of the issues needed to be addressed. Moreover, human factors such as individual skills or demographic factors need to be examined if they have any influence on the performance.

1.1.1 Validation of Lineup

Unlike conventional inference, the lineup protocol uses statistical graphics as test statistics. This conceptual difference leads to defining all the inferential terminologies in a way that allows a non-real test statistic to work. Multiple observers are allowed to evaluate a lineup which gives a certain level of control over Type-I error. To make a decision p-values need to be computed from multiple responses.

To assess how this method performs, the power of the visual test needs to be obtained. To compare power of the visual test it is important to set up a visual test so that a head to head comparison with conventional test can be performed. For this human subject experiment is required. Chapter 2 addressees all these issues in an attempt to validate the lineup protocol. Multiple simulation experiments were done to present the comparisons of power in various scenarios. The paper is accepted by *Journal of the American Statistical Association* for publication.

1.1.2 Human Factors

Human observers have the same role in visual statistical inference as the witness who is asked to pick the suspect from the lineup. It is expected that the human performance would vary depending on the factors such as skill levels, age, gender, education level or geographical location. The individual subject specific power should reflect the possible diversities of human backgrounds.

Other factors such as learning from previous evaluation may have an effect on observer's

choice on the other lineups when an observer evaluates multiple lineups. It could be possible that this learning trend is much different for some people with different demographics. The placement of the data plot in the lineup may have some effect as well since each people may have different way of looking at the lineup while evaluating them. Chapter 3 describes these in details and examines their influence using experimental design. The paper is intended to be submitted to *Sociological Methodology*.

1.1.3 Turk Experiment

One of the challenging part of this research was to get human observers for various experiments done to validate the lineup protocol. We needed observers from a diverse demographics and geographic locations to study the influence of human factors on the power of visual inference. It appears that the most convenient way of doing this is to recruit people from Amazon (2010) Mechanical Turk (MTurk) web site. It is a online workplace where people come to perform simple task and get paid. It is cheap reliable and the results can be obtained very fast. But the tasks that can be designed through MTurk are just too simple to have many control which may be necessary for the human subject experiments on lineups.

Chapter 4 provides a solution to this problem by presenting an alternative way of getting results from MTurk workers. The detailed procedures for designing a human subject experiment are presented, design of an web application is provided for the researchers who may need to recruit people from MTurk to get the lineups evaluated. The web application is now hosted on Iowa State University domain (Majumder, 2013) and multiple experiments were done through this web application. The paper is intended to be submitted to *Journal of Statistical Software*.

1.2 Scope

This thesis forms the building blocks for the new direction of statistical research. It opens up new platform for statistical inference, provides promises to the real problem where no conventional inferential procedure exists. Especially with recent big data problem, visual inference could be very useful because of its non-parametric nature and few assumptions.

Several other followup research works have stemmed from this research. Zhao et al. (2012) conducted an experiment using an eye-tracker to examine which patterns or features participants are cueing on in making their choices while evaluating a lineup. This gives important hints about the effective visual test statistics.

Visual inference is used to study biological applications with high dimension but small sample size data. From simulated data its effectiveness is observed in identifying real separation of data when it exists (Roy Chowdhury et al., 2011). In another published study the power of visual test is used in choosing the best type of plot to convey information for specific applications (Hofmann et al., 2012). Yin et al. (2013) also used visual statistical inference for high-throughput biological data analysis.

A lineup is difficult when it is hard to detect the actual plot and it happens if some of the null plots are very similar to actual data plot. Some distance measures are examined to quantify this similarity of the plots in Roy Chowdhury et al. (2012). For different plot types, the distances can be very different making visual test based on some plot types more powerful than that based on others.

This dissertation work provided a firm background in all these followup studies. Some of them used the web application presented in Chapter 4 to recruit human subjects from Amazon Mechanical Turk to get lineups evaluated. I am an author in the papers described above.

CHAPTER 2. VALIDATION OF VISUAL STATISTICAL INFERENCE, APPLIED TO LINEAR MODELS

A paper accepted by *Journal of the American Statistical Association*

Mahbubul Majumder, Heike Hofmann, Dianne Cook

Abstract

Statistical graphics play a crucial role in exploratory data analysis, model checking and diagnosis. The lineup protocol enables statistical significance testing of visual findings, bridging the gulf between exploratory and inferential statistics. In this paper inferential methods for statistical graphics are developed further by refining the terminology of visual inference, and framing the lineup protocol in a context that allows direct comparison with conventional tests in scenarios when a conventional test exists. This framework is used to compare the performance of the lineup protocol against conventional statistical testing in the scenario of fitting linear models. A human subjects experiment is conducted using simulated data to provide controlled conditions. Results suggest that the lineup protocol performs comparably with the conventional tests, and expectedly out-performs them when data is contaminated, a scenario where assumptions required for performing a conventional test are violated. Surprisingly, visual tests have higher power than the conventional tests when the effect size is large. And, interestingly, there may be some super-visual individuals who yield better performance and power than the conventional test even in the most difficult tasks.

Keywords: statistical graphics, lineup, non-parametric test, data mining, visualization, exploratory data analysis, practical significance, effect size

2.1 Introduction

Statistical graphics nourish the discovery process in data analysis by revealing unexpected things, finding structure that was not previously anticipated, or orthogonally by contrasting prevailing hypotheses. The area of graphics is often associated with exploratory data analysis, which was pioneered by Tukey (1977) and is particularly pertinent in today's data-rich world where discovery during data mining has become an important activity. Graphics are also used in many places where numerical summaries simply do not suffice: model checking, diagnosis, and in the communication of findings.

Several new developments in graphics research have been achieved in recent years. Early studies on evaluating how well statistical plots are perceived and read by the human eye (Cleveland and McGill, 1984), have been repeated and expanded (Simkin and Hastie, 1987; Spence and Lewandowsky, 1991; Heer and Bostock, 2010) with findings supporting the original results. The research by Heer and Bostock (2010) used subjects recruited from Amazon's Mechanical Turk (Amazon, 2010) for their studies. This body of work provides a contemporary framework for evaluating new statistical graphics. In a complementary direction, new research on formalizing statistical graphics with language characteristics makes it easier to abstractly define, compare and contrast data plots. Wilkinson (1999) developed a grammar of graphics that is enhanced by Wickham (2009). These methods provide a mechanism to abstract the way data is mapped to graphical form. Finally, technology advances make it simple and easy for everyone to draw plots of data, and particularly the existence of software systems, such as R (R Development Core Team, 2012), enable making beautiful data graphics that can be tightly coupled with statistical modeling.

However, measuring the strength of patterns seen in plots, and differences in individual perceptual ability, is something that is difficult and perhaps handicaps graphics use among statisticians, where measuring probabilities is of primary importance. This has also been addressed in recent research. Buja et al. (2009) proposes a protocol that allows the testing of discoveries made from statistical graphics. This work represents a major advance for graphics, because it bridges the gulf between conventional statistical inference procedures and exploratory

data analysis. One of the protocols, the lineup, places the actual data plot among a page of plots of null data, and asks a human judge to pick the plot that is different. Figure 2.1 shows an example lineup. Which plot do you think is the most different from the others? (The position of the actual data plot is provided in Section 2.5.1.) Wrapped in a process that mirrors conventional inference, where there is an explicit, a priori, null hypothesis, picking the plot of the data from the null plots represents a rejection of that null hypothesis. The null hypothesis typically derives from the task at hand, or the type of plot being made. The alternative encompasses all possible antitheses, all types of patterns that might be detected in the actual data plot, accounting for all possible deviations from the null without the requirement to specify these ahead of time. The probability of rejection can be quantified, along with Type I, and Type II error, and p-value and power can be defined and estimated.

The protocol has only been informally tested until now. In the work described in this paper, the lineup protocol is compared head to head with the equivalent conventional test. Specifically, the lineup is examined in the context of a linear model setting, where we are determining the importance of including a variable in the model. This is not the envisioned environment for the use of the lineup – actually it is likely the worst case scenario for visual inference. The intended use of lineups is where there is no existing test, and unlikely ever to be any numerical test. The thought is though, that the conventional setting provides a benchmark for how well the lineup protocol works under controlled conditions, and will provide some assurance that they will work in scenarios where there is no benchmark. Testing is done based on a human-subjects experiment using Amazon's Mechanical Turk (Amazon, 2010), using simulation to provide controlled conditions for assessing lineups. The results are compared with those of the conventional test.

The paper is organized as follows. Section 2.2 defines terms as used in visual inference, and describes how to estimate the important quantities from experimental data. The effect of the lineup size and number of observers on the power of the test is discussed in Section 2.3. Section 2.4 focuses on the application of visual inference to linear models. Section 2.5 describes three user studies based on simulation experiments conducted to compare the power of the lineup protocol with the equivalent conventional test and Section 2.6 presents an analysis of

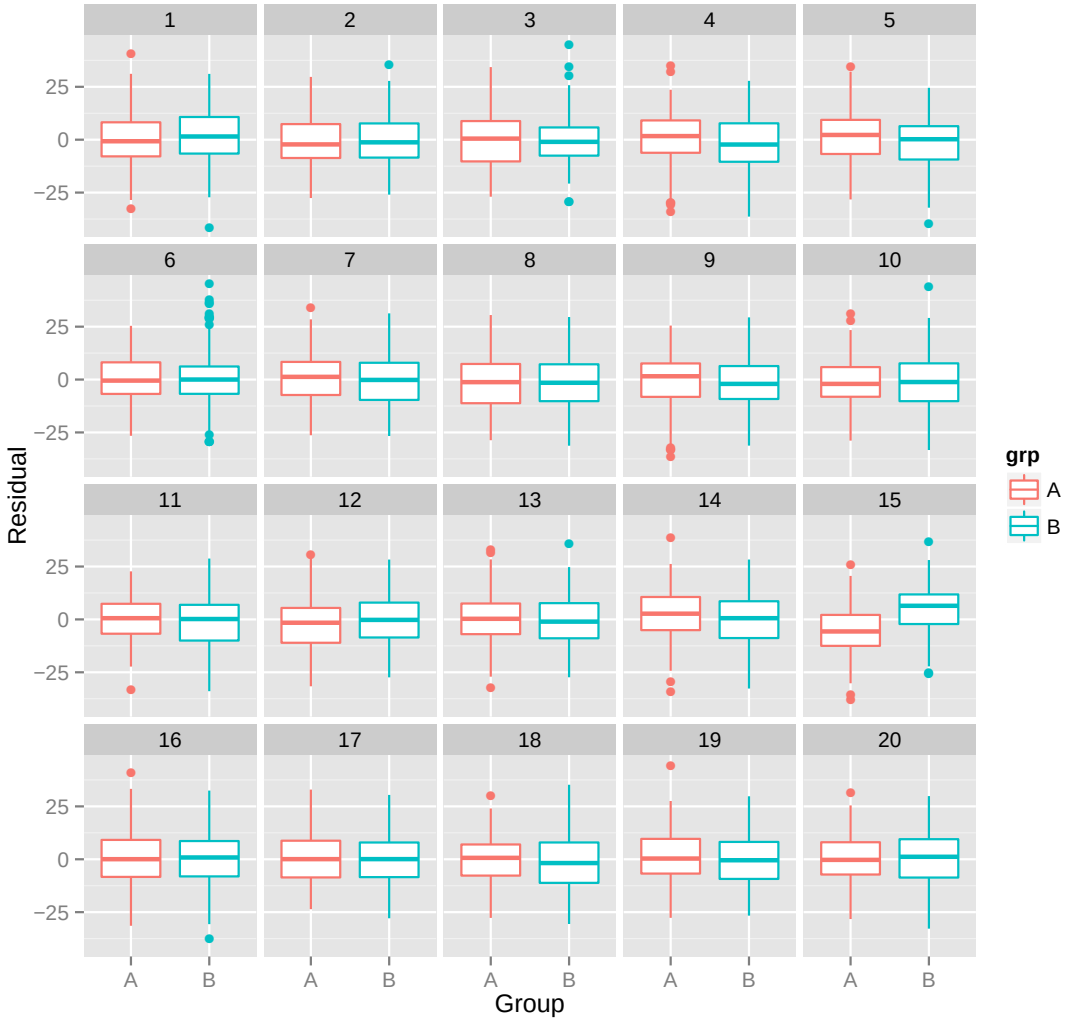


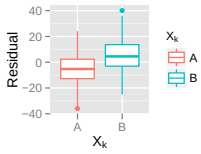
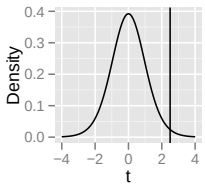
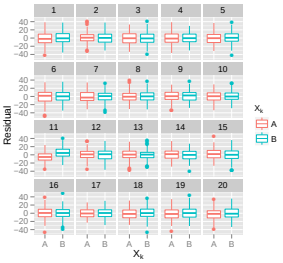
Figure 2.1 Lineup plot ($m = 20$) using side-by-side boxplots for testing $H_0 : \beta_k = 0$. One of these plots is the plot of the actual data, and the remaining are null plots, produced by simulating data from a null model that assumes H_0 is true. Which plot is the most different from the others, in the sense that there is the largest shift or location difference between the boxplots? (The position of the actual data plot is provided in Section 2.5.1.)

the resulting data.

2.2 Definitions and Explanations for Visual Statistical Inference

An illustration of the lineup protocol in relation to conventional hypothesis testing is presented in Table 2.1. Both methods start from the same place, the same set of hypotheses. The conventional test statistic is the t-statistic, where the parameter estimate is divided by its standard error. In the lineup protocol, the test statistic is a plot of the data. Here, side-by-side boxplots are used, because the variable of interest is categorical and takes just two values. In conventional hypothesis testing the value of the test statistic is compared with all possible values of the sampling distribution, the distribution of the statistic if the null hypothesis is true. If it is extreme on this scale then the null hypothesis is rejected. In contrast in visual inference, the plot of the data is compared with a set of plots of samples drawn from the null distribution. If the actual data plot is selected as the most different, then this results in rejection of the null hypothesis.

Table 2.1 Comparison of visual inference with conventional inference.

	Conventional Inference	Lineup Protocol
Hypothesis	$H_0 : \beta = 0$ vs $H_1 : \beta > 0$	$H_0 : \beta = 0$ vs $H_1 : \beta > 0$
	↓	↓
Test statistic	$T(y) = \frac{\hat{\beta}}{se(\hat{\beta})}$	$T(y) =$ 
	↓	↓
Sampling Distribution	$f_{T(y)}(t)$; 	$f_{T(y)}(t)$; 
	↓	↓
Reject H_0 if	actual T is extreme	actual plot is identifiable

In general, we define θ to be a population parameter of interest, with $\theta \in \Theta$, the parameter

space. Any null hypothesis H_0 then partitions the parameter space into Θ_0 and Θ_0^c, with $H_0 : \theta \in \Theta_0$ versus $H_1 : \theta \in \Theta_0^c$. A test statistic, $T(y)$, is a function that maps the sample into a numerical summary, that can be used to test the null hypothesis. The hypothesis test maps the test statistic into $\{0, 1\}$, based on whether $T(y)$ falls into the acceptance region, or the rejection region, respectively. $T(y)$ is assessed relative to null values of this statistic $T(y_0)$, the possible values of T if $\theta \in \Theta$.

For visual inference, unlike in the conventional hypothesis test, the statistic is not a single value, but a graphical representation of the data chosen to display the strength of the parameter of interest, θ. When the alternative hypothesis is true, it is expected that the plot of the actual data, the test statistic, will have visible feature(s) consistent with $\theta \in \Theta_0^c$, and that visual artifacts will not distinguish the test statistic as different when H_1 is not true. We will call a plot with this property a *visual statistic* for θ. More formally,

Definition 2.2.1. *A **visual test statistic**, $T(.)$, is a function of a sample that produces a plot. $T(y)$ maps the actual data to the plot, and we call this the **(actual) data plot**, and $T(y_0)$ maps a sample drawn from the null distribution into the same plot form. These type of plots are called **null plots**.*

Ideally, the visual test statistic is defined and constructed using the grammar of graphics (Wilkinson, 1999; Wickham, 2009), consisting of type and specification of aesthetics, necessary for complete reproducibility. The visual test statistic is compared with values $T(y_0)$ using a lineup, which is defined as:

Definition 2.2.2. *A **lineup** is a layout of m randomly placed visual statistics, consisting of*

- *$m - 1$ statistics, $T(y_0)$, simulated from the model specified by H_0 (null plots) and*
- *the test statistic, $T(y)$, produced by plotting the actual data, possibly arising from H_1.*

The $(m - 1)$ null plots are members of the sampling distribution of the test statistic assuming that the null hypothesis is true. If H_1 is true, we expect this to be reflected as a feature in the test statistic, i.e. the plot of the data, that makes it visually distinguishable from the null plots. A careful visual inspection of the lineup by independent observers follows; observers are

asked to point out the plot most different from the lineup. If the test statistic is identified in the lineup, this is considered as evidence against the null hypothesis. This leads us to a definition for the p-value of a lineup: under the null hypothesis, each observer has a $1/m$ chance of picking the test statistic from the lineup. For K independent observers, let X be the number of observers picking the test statistic from the lineup. Under the null hypothesis $X \sim \text{Binom}_{K,1/m}$, therefore:

Definition 2.2.3. *The p-value of a lineup of size m evaluated by K observers is given as*

$$P(X \geq x) = 1 - Binom_{K,1/m}(x-1) = \sum_{i=x}^{K} \binom{K}{i} \left(\frac{1}{m}\right)^i \left(\frac{m-1}{m}\right)^{K-i}$$

with X defined as above, and x is the number of observers selecting the actual data plot.

Note that for $x = 0$ the p-value becomes, mathematically, equal to 1. It might make more sense from a practical point of view to think of the p-value as being larger than $P(X \geq 1)$ in this situation. By increasing either m or K, the value at a higher precision can be determined. Table 2.2 shows p-values for different numbers of observers for lineups of size $m = 20$.

Table 2.2 Possible p-values for different numbers of observers, K, for fixed size $m = 20$ lineups.

K	x	p-value	K	x	p-value	K	x	p-value	K	x	p-value	K	x	p-value
1	1	0.0500	2	1	0.0975	3	1	0.1426	4	1	0.1855	5	1	0.2262
			2	2	0.0025	3	2	0.0073	4	2	0.0140	5	2	0.0226
						3	3	0.0001	4	3	0.0005	5	3	0.0012
									4	4	< 0.0001	5	4	< 0.0001

Definition 2.2.4. *The **visual test**, V_θ of size m and significance level α, is defined to*

- ***Reject** H_0 if out of K observers at least x_α correctly identify the actual data plot, and*
- ***Fail to reject** H_0 otherwise.*

where x_α is such that $P(X \geq x_\alpha | H_0) \leq \alpha$.

Associated with any test there is the risk of Type I or II errors, which for visual inference are defined as follows:

Definition 2.2.5. *The **Type I error** associated with visual test V_θ is the probability of rejecting H_0 when it is true; the probability for that is $P(X \geq x_\alpha)$, which is controlled by α. The **Type II error** is the probability of failing to identify the actual data plot, when H_0 is not true, $P(X < x_\alpha)$.*

Because X takes only discrete values we can not always control exactly for α. For example, when there is only one observer, $1/m$ is the minimal value at which we can set α. It can be set to be smaller, even arbitrarily small, by increasing K, the number of observers. Type II error is harder to calculate, as is usually the case. In visual inference, individual abilities need to be incorporated to calculate Type II error. Here, we need to estimate the probability that an observer sees the actual data plot as different, when it really is different. This involves understanding the individual's visual skills. Thus, let X_i be a binary random variable with $X_i = 1$, if individual i $(= 1, \ldots, K)$ identifies the actual data plot from the lineup, and $X_i = 0$ otherwise. Let p_i be the probability that individual i picks out the actual data plot. If all individuals have the same ability, with the probability, p, for picking out the actual data plot, then $X = \sum_i X_i$ has distribution $\text{Binom}_{K,p}$, and we can estimate p by $\hat{p} = x/K$, where x is the number of observers (out of K), who pick out the actual data plot.

If there is evidence for individual skills influencing the probability p_i, then $X_i \sim Binom_{1,p_i}$ and X is a sum of independent Bernoulli random variables with different success rates p_i. This makes the distribution of X a Poisson-Binomial by definition (see Butler and Stephens (1993) for details). Ways to estimate p_i will be discussed in the following sections.

Definition 2.2.6. *The **power** of a visual test, V_θ, is defined as the probability to reject the null hypothesis for a given parameter value θ:*

$$Power_V(\theta) = Pr(Reject\ H_0 \mid \theta)$$

An important difference between conventional and visual testing is that lineups will depend on observers' evaluation. Thus X, the number of observers who identify the actual data plot from the lineup, affects the estimation of power and the power is estimated by

$$\widehat{Power}_V(\theta) = Power_{V,K}(\theta) = 1 - F_{X,\theta}(x_\alpha - 1).$$

Here $F_{X,\theta}$ is the distribution of X and x_α is such that $P(X \geq x_\alpha) \leq \alpha$. Note that the distribution F_X depends on which hypothesis is true: under the null hypothesis, $X \sim Binom_{K,1/m}$, leading to:

$$Power_V(\theta, K) = 1 - Binom_{K,1/m}(x_\alpha - 1).$$

If the alternative hypothesis is true, with a fixed parameter value θ, we can assume that an individual's probability to identify the data plot depends on the parameter value, and $X_i \sim Binom_{1,p_i(\theta)}$. Assessing an individual's skill to identify the actual data plot will require that an individual evaluates multiple lineups.

Power is an important consideration in deciding which test to use for solving a problem. Here we use it to compare the performance of the visual test with the conventional test, but in practice for visual inference it will mostly be important in choosing plots to use. Analysts typically have a choice of plots to make, and a myriad of possible options such as reference grids, for any particular purpose. This is akin to different choices of statistics in conventional hypothesis testing, for example, mean, median, or trimmed mean. One is typically better than another. For two different visual test statistics of the same actual data, one is considered to be better, if $T(y)$ is more easily distinguishable to the observer. Power is typically used to measure this characteristic of a test.

2.3 Effect of Observer Skills and Lineup Size

2.3.1 Subject-specific abilities

Suppose each of K independent observers gives evaluations on multiple lineups, and responses are considered to be binary random variable, $X_{\ell i} \sim Binom_{1,p_{\ell i}}$, where $X_{\ell i} = 1$, if subject i correctly identifies the actual data plot on lineup ℓ, $1 \leq \ell \leq L$, and 0 otherwise. A mixed effects logistic regression model is used for $P(X_{\ell i} = 1) = p_{\ell i} = E(X_{\ell i})$, accommodating both for different abilities of observers as well as differences in the difficulty of lineups.

The model can be fit as:

$$g(p_{\ell i}) = W_{\ell i}\delta + Z_{\ell i}\tau_{\ell i}, \tag{2.1}$$

where $g(.)$ denotes the *logit* link function $g(\pi) = \log(\pi) - \log(1-\pi); 0 \leq \pi \leq 1$. W is a design matrix of covariates corresponding to specifics of lineup ℓ and subject i, and δ is the vector of corresponding parameters. Covariates could include demographic information of individuals, such as age, gender, education level etc., as well lineup-specific elements, e.g. effect size or difficulty level. $Z_{\ell i}$, $1 \leq i \leq K$, $1 \leq \ell \leq L$, is a design matrix corresponding to random effects specific to individual i and lineup ℓ; and τ is a vector of independent normally distributed random variables $\tau_{\ell i}$ with variance matrix $\sigma_\tau I_{KL \times KL}$. τ will usually include a component incorporating an individual's ability or skill to evaluate lineups. Note that $\tau_{\ell i}$ usually only includes a partial interaction; for a full interaction of subjects' skills and lineup-specific difficulty we would need replicates of the same subject evaluating the same lineup, which in practice is not feasible without losing independence.

The inverse *logit* link function, $g^{-1}(.)$, from Equation 2.1 leads to the estimate of the subject and the lineup specific probability of successful evaluation by a single observer as

$$\hat{p}_{\ell i} = g^{-1}(W_{\ell i}\hat{\delta} + Z_{\ell i}\hat{\tau}_{\ell i}). \tag{2.2}$$

2.3.2 Lineup size, m

The finite number $m-1$ of representatives of the null distribution used as comparison against the test statistic, is a major difference between visual inference and the conventional testing. The choice of m has an obvious impact on the test.

The following properties can only be derived for the situation of a fully parameterized simulation study, as conducted in this paper. They allow for a direct comparison of lineup tests against the conventional counterparts, and also to identify properties relevant for a quality assessment of lineups when they are used in practical settings. Two assumptions are critical:

1. the plot setup is structured in a way that makes it possible for an observer to identify a deviation from the null hypothesis,

2. an observer is able to identify the plot with the strongest 'signal' (or deviation from H_0) from a lineup.

Evidence in support of the second assumption will be seen in the data from the study discussed in Section 2.5, the degree to which the first assumption is fulfilled is reflected by the power of a lineup. The better suited a design is for a particular task, the higher its power will be.

In order to compare the power of conventional and visual tests side-by-side, it is necessary to assume that we are in the controlled environment of a simulation with tests corresponding to a known parameter value $\theta \in R$ and associated distribution function F_t of the test statistic.

Lemma 2.3.1. *Suppose $F_{|t|}(.)$ is the distribution function of an absolute value of t, the conventional test statistic. Suppose the associated test statistic is observed as t_{obs} with p-value p_D.*

The probability of picking the data plot from a lineup depends on the size m of the lineup and the strength of the signal in the data plot. Under the above assumptions, the probability is expressed as:

$$P(p_D < p_0) = E\left[(1 - p_D)^{m-1}\right]$$

where p_D is the p-value associated with the data in the test statistic, and p_0 is the minimum of all p-values in the data going into null plots.

Proof. The proof and the details of the lemma are attached in the supplementary documents. □

The above lemma allows two immediate conclusions for the use of lineups. The probability that the observer correctly identifies the data plot is closely connected to the size of the lineup m, since the right hand side of the above equation decreases for larger m, the probability of correctly identifying the actual data plot decreases with m. Further we see, that the rate of this decrease depends strongly on the distribution of p_D – if the density of p_D is very right skewed, the expectation term on right hand side will be large and less affected by an increase in m. This can also be seen in Figure 2.2, which illustrates lemma 2.3.1. Figure 2.2 shows the probability of picking the actual data plot for lineups of different size: as m increases we have

an increased probability to observe a more highly structured null plot by chance. It can also be seen that for a p-value, p_D, of about 0.15 for the data plot, the signal in the plot is so weak, that it cannot be distinguished from null plots in a lineup of size $m = 20$.

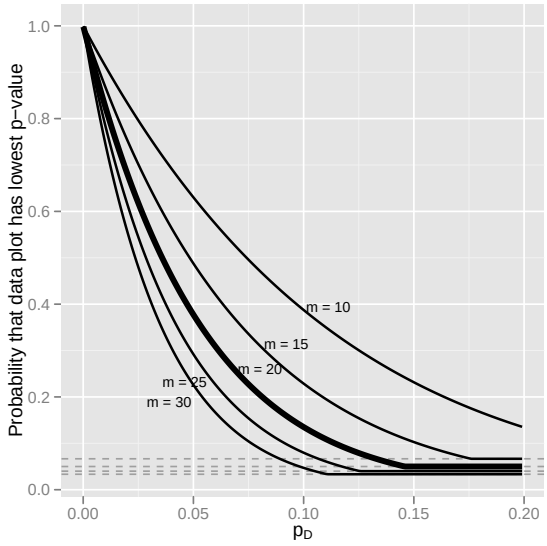


Figure 2.2 Probability that the data plot has the smallest probability in a lineup of size m. With increasing p-value the probability drops – when it reaches $1/m$ a horizontal line is drawn to emphasize insufficient sensitivity of the test due to the lineup size.

2.4 Application to Linear Models

To make these concepts more concrete consider how this would operate in the linear models setting. Consider a linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i1} X_{i2} + ... + \epsilon_i \tag{2.3}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, 2, .., n$. The covariates $(X_j, j = 1, .., p)$ can be continuous or discrete.

In this setting, there are many established graphics that are used to evaluate and diagnose the fit of a regression model (e.g. Cook and Weisberg 1999). Table 2.3 lists several common hypotheses related to the regression setting, and commonly used statistical plots that might be used as corresponding visual test statistics. For example, to examine the effect of variable X_j on Y, we would plot residuals obtained from fitting the model without X_j against X_j or

for a single covariate we may plot Y against X_j (cases 1-4 in Table 2.3). To assess whether the assumption of linearity is appropriate we would draw a plot of residuals against fitted values (case 5 in Table 2.3). For the purpose of comparing visual against conventional inference, we focus on cases 2 and 3, with a continuous and categorical explanatory variable, respectively.

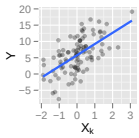
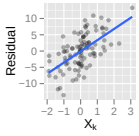
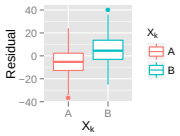
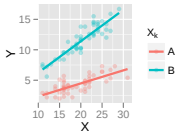
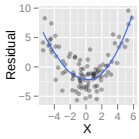
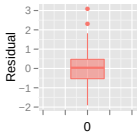
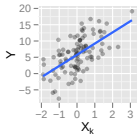
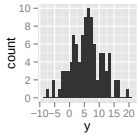
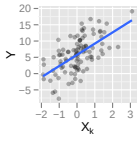
Suppose X_k is a categorical variable with two levels, and we test the hypothesis $H_0 : \beta_k = 0$ vs $H_1 : \beta_k \neq 0$. If the responses for the two levels of the categorical variable X_k in the model are different, the residuals from fitting the null model should show a significant difference between the two groups. For a visual test, we draw boxplots of the residuals conditioned on the two levels of X_k. If $\beta_k \neq 0$ the boxplots should show a vertical displacement.

The conventional test in this scenario uses $T = \hat{\beta}_k / se(\hat{\beta}_k)$ and rejects the null hypothesis, if T is extreme on the scale of a t distribution with $n - p$ degrees of freedom. It forms the benchmark upon which we evaluate the visual test. To calculate what we might expect for the power of the visual test, under perfect conditions, first assume that the observer is able to pick the plot with the smallest p-value from a lineup plot. This leads to the decision to reject H_0 when $p_D < p_0$, where p_D is the conventional p-value as details given in Lemma 2.3.1. Thus the expected probability to reject by a single observer ($K = 1$) in this scenario is

$$p(\beta) = Pr(p_D < p_0) \quad \text{for} \quad \beta \neq 0 \tag{2.4}$$

Figure 2.3 shows the power of the conventional test in comparison to the expected power of the visual test for different K (number of observers), obtained using $p(\beta)$ from Equation 2.4. Notice that the expected power of the visual test exceeds the power of the conventional test as K increases, and when β gets larger. Conversely, visual power is below conventional power for parameter values close to the null hypothesis. This is even more pronounced for large number of observers. At the same time, the point of intersection between visual and conventional power approaches the value of the null hypothesis as the number of observers approaches infinity, leading to an asymptotically perfect power curve of zero in the null hypothesis and one for any alternative value. We observe this dichotomy of visual power in power estimates based on the data collected from user experiments, too. It features prominently in figure 2.6.

Table 2.3 Visual test statistics for testing hypotheses related to the model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_1 X_{i2} + \beta_3 X_{i1} X_{i2} + ... + \epsilon_i$

Case	Null Hypothesis	Statistic	Test Statistic	Description
1	$H_0 : \beta_0 = 0$	Scatter plot		Scatter plot with least square line overlaid. For null plots we simulate data from fitted null model.
2	$H_0 : \beta_k = 0$	Residual plot		Residual vs X_k plots. For null plots we simulate data from normal with mean 0 variance $\hat{\sigma}^2$.
3	$H_0 : \beta_k = 0$ (for binary X_k)	Box plot		Box plot of residuals grouped by category of X_k. For null plots we simulate data from normal with mean 0 variance $\hat{\sigma}^2$.
4	$H_0 : \beta_k = 0$ (interaction of continuous and binary X_k)	Scatter plot		Scatter plot with least square lines of each category overlaid. For null plots we simulate data from fitted null model.
5	$H_0 : X$ Linear	Residual Plot		Residual vs predictor plots with loess smoother overlaid. For null plots we simulate residual data from normal with mean 0 variance $\hat{\sigma}^2$.
6	$H_0 : \sigma^2 = \sigma_0^2$	Box plot		Box plot of standardized residual divided by σ_0^2. For null plots we simulate data from standard normal.
7	$H_0 : \rho_{X,Y\|Z} = \rho$	Scatter Plot		Scatter plot of residuals obtained by fitting partial regression. For null plots we simulate data (mean 0 and variance 1) with specific correlation ρ.
8	H_0 : Model Fits	Histogram		Histogram of the response data. For null plots we simulate data from fitted model.
9	For $p = 1$ only: $H_0 : \rho_{X,Y} = \rho$	Scatter plot		Scatter plot with least square line overlaid. For null plots we simulate data with correlation ρ.

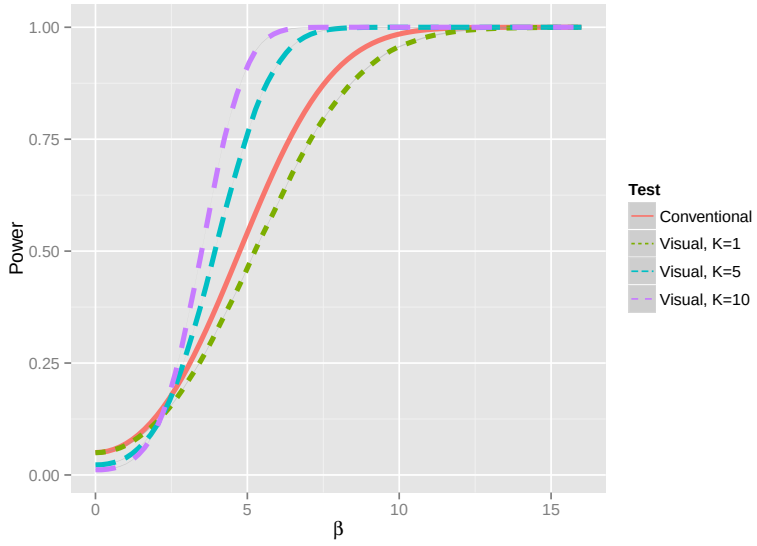


Figure 2.3 Comparison of the expected power of a visual test of size $m = 20$ for different K (number of observers) with the power of the conventional test, for $n = 100$ and $\sigma = 12$.

2.5 Human Subjects Experiments with Simulated Data

Three experiments were conducted to evaluate the effectiveness of the lineup protocol relative to the equivalent test statistic used in the regression setting. The first two experiments have ideal scenarios for conventional testing, where we would not expect the lineup protocol to do better than the conventional test. The third experiment is a scenario where assumptions required for the conventional test are violated, and we would expect the lineup protocol to outperform the conventional test. (Data and lineups used in the experiments are available in the supplementary material.)

After many small pilot studies with local personnel, it was clear that some care was needed to set up the human subjects experiments. It was best for a observer or a subject to see a block of 10 lineups with varying difficulty, with a reasonable number of "easy" lineups. The explanations about each experiment (below) includes an explanation of how the lineups were sampled and provided to the subjects.

Participants for all the experiments were recruited through Amazon's online web service, Mechanical Turk (Amazon, 2010). A summary of the data obtained for all three experiments

are shown in Table A.2. Participants were asked to select the plot they think best matched the question given, provide a reason for their choice, and say how confident they are in their choice. Gender, age, education and geographic location of each participant are also collected. For each of the experiments one of the lineups was used as a test plot (easy plot) which everyone should get correct, so that a measure of the quality of the subjects effort could be made. Note that no participant was shown the same lineup twice.

2.5.1 Discrete covariate

The experiment is designed to study the ability of human subjects to detect the effect of a single categorical variable X_2 (corresponding to parameter β_2) in a two variable ($p = 2$) regression model (Equation 2.3). Data is simulated using a range of values of β_2 or slopes as shown in Table 2.4, two different sample sizes ($n = 100, 300$) and two standard deviations of the error ($\sigma = 5, 12$). The range of β_2 values was chosen so that estimates of the power would produce reasonably continuous power curves, comparable to that calculated for the theoretical conventional test. Values were fixed for other regression parameters, $\beta_0 = 5$, $\beta_1 = 15$, and the values for X_1 were randomly generated from a Poisson ($\lambda = 30$) distribution, which is almost Gaussian. Three data sets were generated for each of the parameter values shown in Table 2.4 resulting in 60 different "actual data sets", and thus, 60 different lineups. For each lineup, the null model was fit to the actual data set to obtain residuals and parameter estimates. The actual data plot was drawn as side-by-side boxplots of the residuals (Table 2.3, case 3). The 19 null data sets were generated by simulating from $N(0, \hat{\sigma}^2)$, and plotted in the same way. The actual data plot was randomly placed among these null data plots to produce the lineup. Figure 2.1 is an example of one of these lineups. It was generated for n=300, β_2=10 and σ=12. The actual data plot location is $(4^2 - 1)$. For this lineup, 15 out of 16 observers picked the actual data plot.

The number of evaluations required for each lineup to provide reasonable estimates of the proportion correct ($\hat{p}$) is determined by the variance of the number of correct evaluations. Suppose γ denotes the conventional test power for each parameter combination shown in Table 2.4. Since the expected power of visual inference is very close to the power of conventional test

(Figure 2.3 with $K = 1$) we consider $\gamma = p$. For a given proportion γ it is desired to have a margin of error (ME) less than or equal to 0.05. Thus we have $ME = 1.96\sqrt{\gamma(1-\gamma)/n_\gamma} \leq 0.05$ which gives us the estimation of minimum number of evaluations

$$n_\gamma \geq \frac{\gamma(1-\gamma)}{(0.05/1.96)^2}.$$

Each subject viewed at least 10 lineups with the option to evaluate more. Depending on the parameter combinations we group the lineups in different difficulty levels as easy, medium, hard and mixed (actual numbers are given in the supplementary material). For each difficulty level a specific number of lineups was randomly picked for evaluation. This number is chosen so that total number of evaluations for each lineup for that group exceed the threshold n_γ. To satisfy this plan we needed to recruit at least 300 subjects.

Table 2.4 Combination of parameter values, β_2, n and σ, used for the simulation experiments.

Sample size (n)	Error SD (σ)	Slope (β) Experiment 1 Discrete covariate	Experiment 2 Continuous covariate	Experiment 3 Contaminated data
100	5	0, 1, 3, 5, 8	0.25, 0.75, 1.25, 1.75, 2.75	0.1, 0.4, 0.75, 1.25, 1.5, 2.25
	12	1, 3, 8, 10, 16	0.5, 1.5, 3.5, 4.5, 6	
300	5	0, 1, 2, 3, 5	0.1, 0.4, 0.7, 1, 1.5	
	12	1, 3, 5, 7, 10	0, 0.8, 1.75, 2.3, 3.5	

2.5.2 Continuous covariate

This experiment is very similar to the previous one, except that there is a single continuous covariate and no second covariate (Equation 2.3 with $p = 1$), following the test in Table 2.3, case 2. Data is simulated with two sample sizes ($n = 100, 300$), two standard deviations of the error ($\sigma = 5, 12$), a variety of slopes (β), as given in Table 2.4. We arbitrarily set $\beta_0 = 6$ and values for X_1 are simulated from $N(0, 1)$. For each combination of parameters, at least three different actual data sets are produced, yielding a total of 70 lineups.

The actual data plot is generated by making a scatterplot of Y vs X_1 with the least squares regression line overlaid. To produce the null plots in the lineup null data was simulated from $N(X\hat{\beta}, \hat{\sigma}^2)$ and plotted using the same scatterplot method as the actual data. To select 10 lineups for a subject, each combination of sample size (n) and error SD (σ) is given a difficulty

value based on the slope (β) parameters. For the smallest slopes the difficulty is 4 (hardest) and for the largest slopes the difficulty is 0 (easiest). Figure 2.4 shows an example lineup for this experiment from difficulty level 4. This lineup is generated using a sample size (n) of 100, slope (β) of 1.25 and and error SD (σ) of 5. The actual data plot location is $(2^2 + 1)$. None of the 65 observers picked the actual plot while 46 observers picked plot 18 which has the lowest p-value among all the plots in this lineup.

For each combination of sample size and standard deviation, each participant is given five randomly selected lineups, one of each difficulty level. Another set of four lineups is chosen from a second tier of selected combinations of sample size and standard deviation, with difficulty levels 0 to 3. A last lineup was randomly selected from a set of lineups with difficulty level 0. The order in which the lineups are shown to participants is randomized.

2.5.3 Contaminated data

The first two simulation experiments use data generated under a normal error model, satisfying the conditions for conventional test procedures. In these situations there exists a test, and there would, in general, be no need to use visual inference. The simulation is conducted in the hope that the visual test procedure, will at least compare favorably with the conventional test – without any ambition of performing equally well. This third simulation is closer to the mark for the purpose of visual inference. The assumptions for the conventional test are violated by contaminating the data. The contamination makes the estimated slopes effectively 0, yet the true value of slope parameter is not. The data is generated from the following model:

$$Y_i = \begin{cases} \alpha + \beta X_i + \epsilon_i & X_i \sim N(0,1) \quad i = 1, ..., n \\ \lambda + \eta_i & X_i \sim N(\mu, 1/3) \quad i = 1, ..., n_c \end{cases}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma)$, $\eta_i \overset{iid}{\sim} N(0, \sigma/3)$ and $\mu = -1.75$. n_c is the size of the contaminated data. For the experiment we consider $n = 100$ and $n_c = 15$ producing actual data with 115 points. Further, $\alpha = 0$, $\lambda = 10$, and σ is chosen to be approximately 3.5, so that error standard deviation across both groups of the data is 5. A linear model (Equation 2.3 with $p = 1$ and intercept $\beta_0 = 0$) is fit to the contaminated data. This experiment follows the test in Table 2.3,

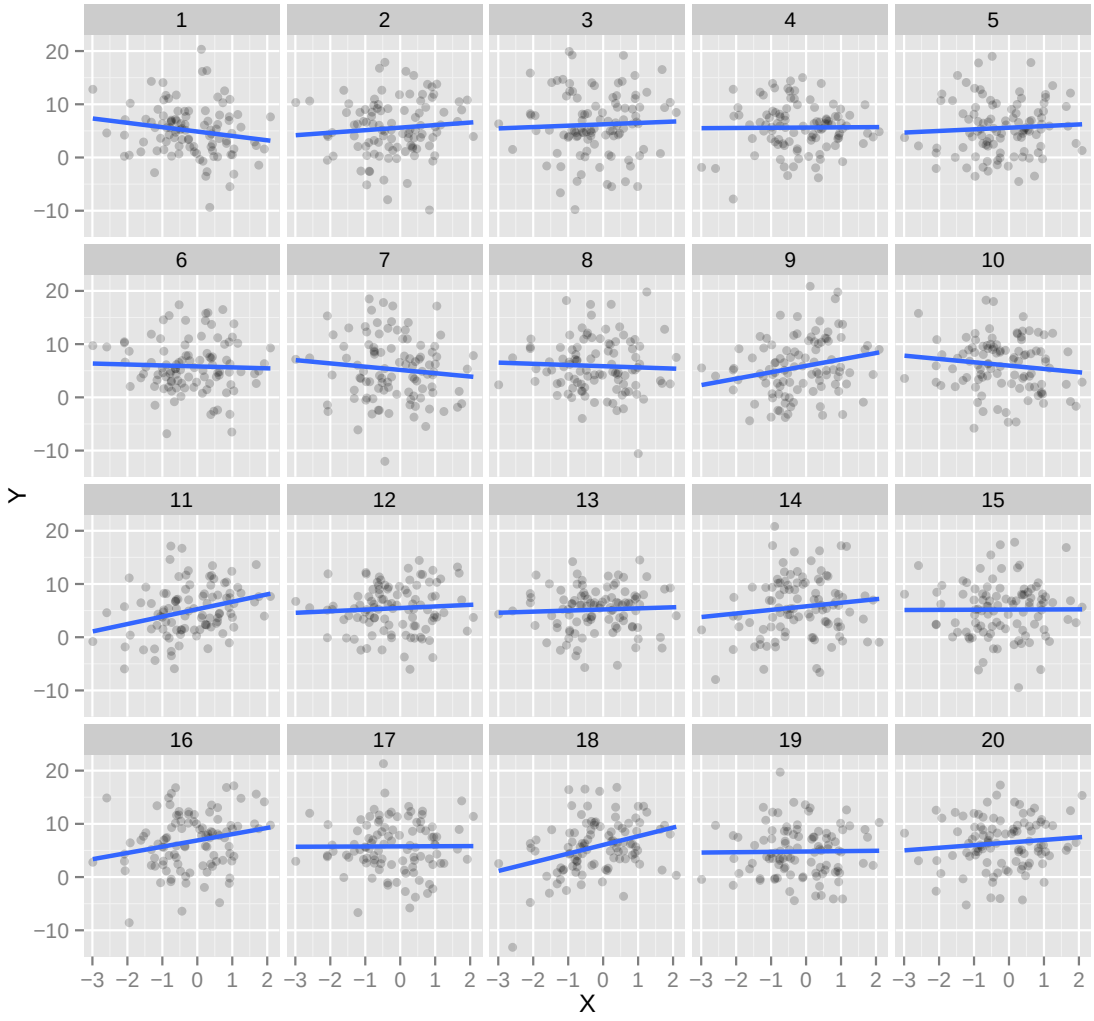


Figure 2.4 Lineup plot ($m = 20$) using scatter plots for testing $H_0 : \beta_k = 0$ where covariate X_k is continuous. One of these plots is the plot of the actual data, and the remaining are null plots, produced by simulating data from a null model that assumes H_0 is true. Which plot is the most different from the others, in the sense that there is the steepest slope? (The position of the actual data plot is provided in Section 2.5.2.)

case 2. The actual data plot shows a scatterplot of the residuals vs X_1, and the null plots are scatterplots of null data generated by plotting simulated residuals from $N(0, \hat{\sigma}^2)$ against X_1.

Experiment three consists of a total of 30 lineups, made up of five replicates for each of the six slopes as shown in Table 2.4. We use the slope directly as a measure of difficulty, with difficulty = 0 for the largest slope and difficulty = 5 for the smallest slope. Subjects were exposed to a total of ten lineups, with two lineups from each of the difficulty levels 0 through 3, and one lineup each from levels 4 and 5.

An example lineup for slope $\beta = 0.4$ is shown in Figure 2.5. Can you pick which plot is different? The actual data plot location is $(3^2 - 2^3)$ and 13 out of 31 observers picked the actual plot.

2.6 Results

2.6.1 Data Cleaning

Amazon Mechanical Turk workers are paid for their efforts, not substantially, but on the scale of the minimum wage in the USA. Some workers will try to maximize their earnings for minimum effort, which can affect the results from the data. For example, some workers may simply randomly pick a plot, without actively examining the plots in the lineup. For the purpose of identifying these participants and cleaning the data, we use one of the very easy lineups that everybody was exposed to as a *reference lineup* and take action based on a subject's answer to this reference: if the subject failed to identify the actual data plot on the reference lineup, we remove all of this subject's data from the analysis. If the answer on the reference lineup is correct, we remove the answer for this lineup from the analysis, but keep all of the remaining answers. Table A.2 tabulates the number of subjects, genders and lineups evaluated after applying the data screening procedure.

2.6.2 Model fitting

For each parameter combination, *effect* E is derived as $E = \sqrt{n} \cdot \beta / \sigma$. The model in Equation 2.1 is fit using E as the only fixed effect covariate without intercept, i.e. $W_{\ell i} = E_{\ell i}$. Instead of

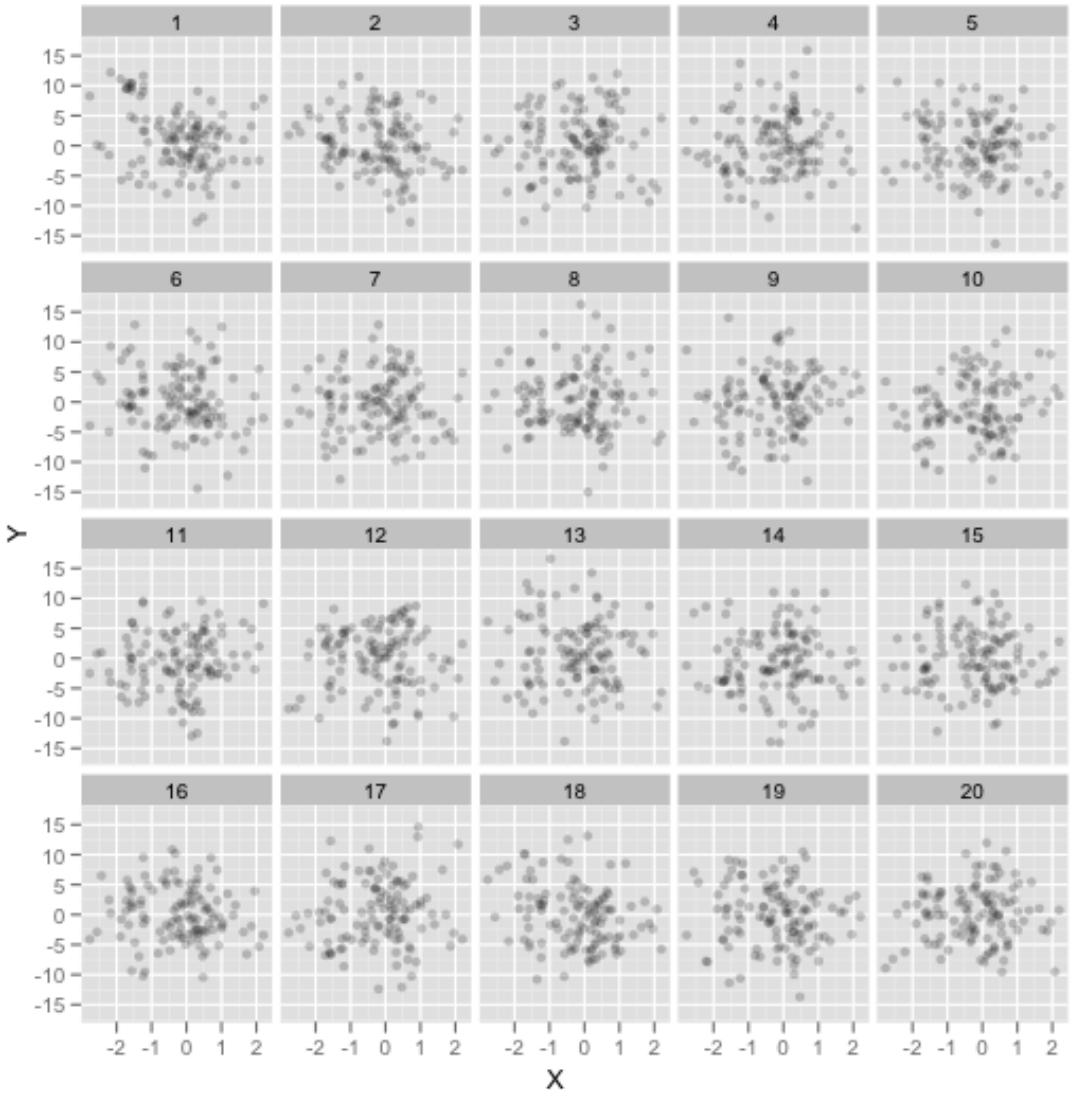


Figure 2.5 Lineup plot ($m = 20$) using scatter plots for testing $H_0 : \beta_k = 0$ where covariate X_k is continuous but the inclusion of some contamination with the data spoils the normality assumption of error structure. One of these plots is the plot of the actual data, and the remaining are null plots, produced by simulating data from a null model that assumes H_0 is true. Which plot is the most different from the others, in the sense that there is the steepest slope? (The position of the actual data plot is provided in Section 2.5.3.)

Table 2.5 Number of subjects, gender, total lineups seen and distinct lineups for all three experimental data sets. Note that in some of the lineups the number of male and female participants does not add up to the total number of participants due to missing demographic information.

Experiment	Subject	Male	Female	Responses	Lineup
1	239	121	107	2249	60
2	351	185	164	3636	70
3	155	103	52	1511	29

fitting an intercept, we make use of a fixed offset of $\log(0.05/0.95)$ so that the estimated power has a fixed lower limit at 0.05 (Type-I error) when $E = 0$. Different skill levels of subjects are accounted for by allowing subject-specific random slopes for effect (E).

For experiment 3 we do fit intercepts: both a fixed and subject-specific random effects, since forcing power to be fixed at 0.05 for $E = 0$ is not required by the experimental design.

For computation we use package `lme4` (Bates et al., 2011) and software R 2.15.0 (R Development Core Team, 2012). p-value calculations are based on asymptotic normality.

Table 2.6 shows the parameter estimates of the mixed effects model of the subject-specific variation. The fixed effects estimates indicate that for all experiments the proportion of correct responses increases as the effect increases. This effect is less pronounced for experiment 3. The subject-specific variability is smaller for experiment 1, and relatively large for experiment 3.

Table 2.6 Parameter estimates of model in Equation 2.1. Estimates are highly significant with p-value < 0.0001 for all three experiment data.

	Fixed effect		Random effect
Experiment	Estimate	Std. error	Variance
1	0.39	0.0094	0.0080
2	1.21	0.0197	0.0443
3	0.59 (Intercept)	0.1668	1.9917
	0.21 (Slope)	0.0511	0.0245
	-0.78 (correlation)		

2.6.3 Power comparison

Figure 2.6 shows an overview of estimated power against effect for the three experiments. Responses from each experiment are summarized by effect size and represented as dots, with size indicating the number of responses. A loess fit to the data gives an estimate of the observed proportion correct $\hat{p}(E)$ for different effect sizes, with grey bands indicating simultaneous bootstrap confidence bands (Buja and Rolke, 2011). $\hat{p}(E)$ is considered to be the power for $K = 1$ and it is used to obtain power for $K = 5$. For comparison, the dashed lines show the corresponding power curves of the conventional tests. It is encouraging to see that visual inference mirrors the power vs effect relationship of conventional testing, in experiments 1 and 2. In experiment 3 the power of the visual test exceeds that for the conventional test, as expected. For larger values of K estimated power exceeds the power of conventional test. Note that for effect $E = 0$, the power is close to 0.05 (Type-I error) for both experiments 1 and 2, making the fixed offset a reasonable assumption.

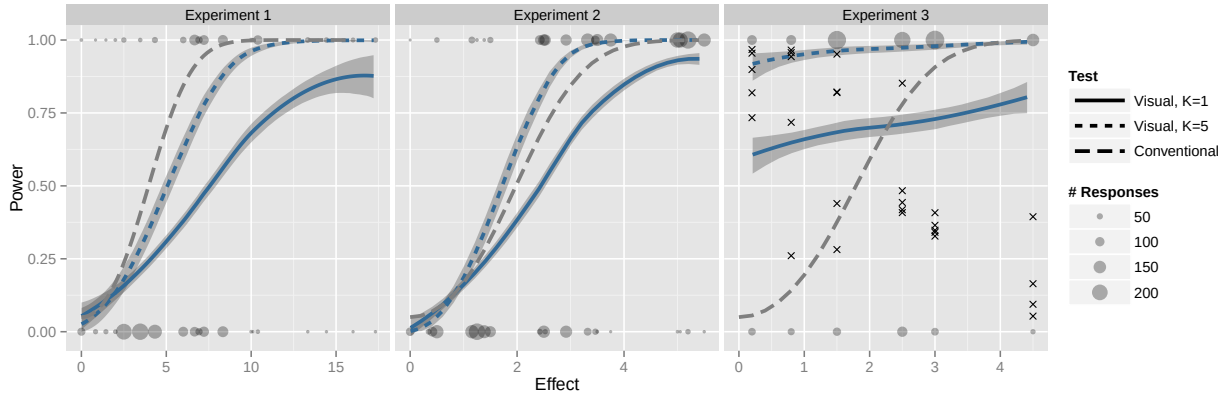


Figure 2.6 Power in comparison to effect for the three experiments. Points indicate subject responses, with size indicating count. Responses are 1 and 0 depending on the success or failure respectively to identify the actual plot in the lineup. The loess curve (continuous line) estimates the observed proportion correct (power for $K = 1$), and surrounding bands show simultaneous bootstrap confidence band. Observed proportion is used to obtain power for $K = 5$. Conventional test power is drawn as a dashed line. For experiment 3, conventional power is based on the slopes of the non-contaminated part of the data. Power of the conventional test for contaminated data is shown by cross marks.

Results for experiment 3 are quite different. This is the situation where we expect to see the

potential of visual inference, and indeed we do: the power of visual inference is always high, and much higher than the conventional test at small effect sizes. There is no actual conventional power in this situation, because assumptions are violated. The dashed line shows conventional power based on uncontaminated data, whereas the cross marks show effective power based on the coefficient estimated from the contaminated data.

Results of experiment 3 are curious insofar, as power of the visual test is largely independent of effect size. However, these results are based on correct identification of the actual data plot, regardless of reason. Although subjects were asked to select the plot that exhibited the highest association between the two variables, they might have cued in on the cluster of contaminated data. This will be explored further in Section 2.6.8.

2.6.4 Subject-specific variation

Subject-specific proportion correct $\hat{p}_i(E)$ is obtained using Equation 2.2 and it is used to obtain power for $K = 5$. Figure 2.7 shows power curves for both the overall experiment and subject-specific variations. The thick continuous line shows overall estimated power, the thinner lines correspond to subject-specific power curves. For comparison, the dashed lines show power curves of the conventional test. Subject-specific power is quite different between the three experiments. In experiment 2 subjects performed similarly, and substantially better than the conventional test. In experiment 1 there is more variability between subjects, with some doing better than the conventional test on large effects. In experiment 3 there is the most subject-specific variation. Some subjects performed substantially better than the conventional test, and on average the visual test was better.

2.6.5 Estimating the p-value in the real world

In the real setting, where visual inference is to be useful, there will be no conventional test p-values. Assessing the strength of perceived structure is a critical component of visual inference. In experiments 1 and 2, there is a p-value associated with the actual data plot in each lineup. As the p-value increases the proportion of correct responses falls (Figure 2.8), which is evidence of direct association between proportion of correct responses and conventional test

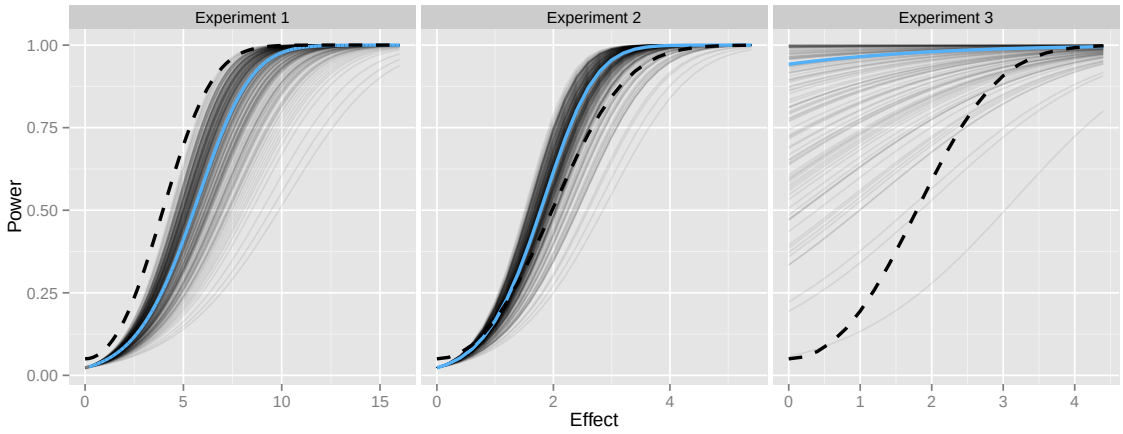


Figure 2.7 Subject-specific power for $K = 5$ obtained using the subject-specific proportion correct estimated from model 2.1. The corresponding power curve for conventional test (dashed line) is shown for comparison. The overall estimated average power curve is shown (light blue).

p-values. For p-values larger than 0.15 it is very uncommon for subjects to correctly identify the actual data plot in the lineup.

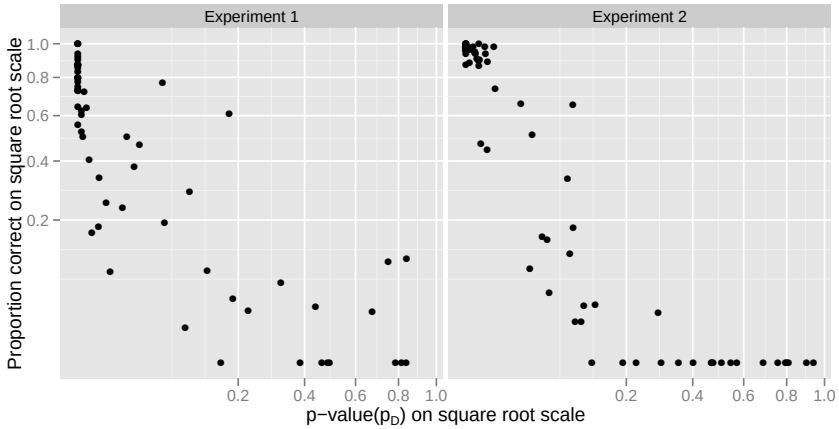


Figure 2.8 Proportion of correct responses decreases rapidly with increasing p-values. For p-values above 0.15 it becomes very unlikely that observers identify the actual plot. The theoretical justification of this is shown in Figure 2.2.

From the experimental data the visual p-values are estimated based on Definition 2.2.3. Figure 2.9 displays resulting estimates for each lineup against the conventional p-value. The pattern of visual p-values is interesting: for small p-values the visual estimates tend to be very small, while lineups with larger p-values result in very large visual estimate, giving a clear

indication to reject H_0 or not. This is why we do not see lot of visual p-values between 0.05 and 0.8 especially for experiment 2. This guides the researcher to make decision confidently while conventional tests with marginal p-values make the decision whether to reject or not harder. For visual tests this is not common.

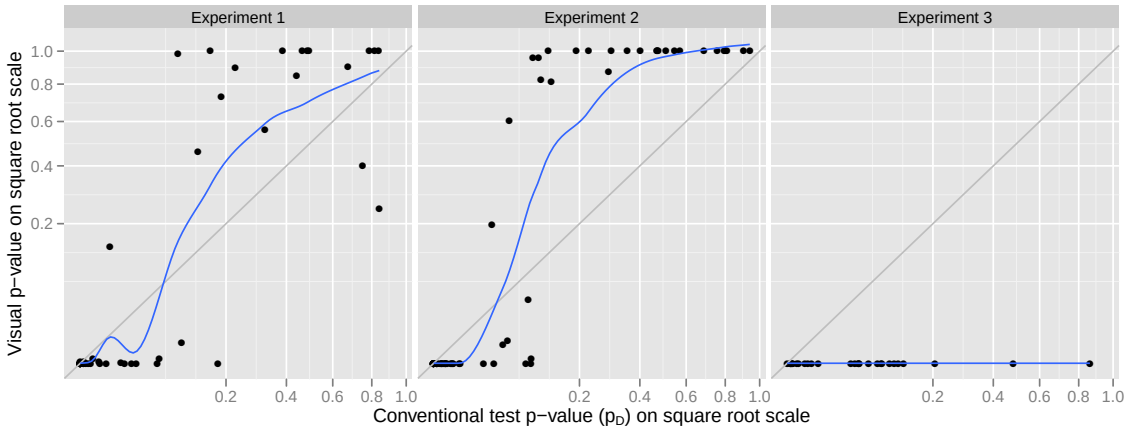


Figure 2.9 Conventional test p-value (p_D) vs visual p-value obtained from the definition . Values are shown on square root scale.

For experiment 3, we see that the visual p values are very small no matter what the conventional p-values are. This is expected as the conventional test loses its power to reject H_0 even when the alternative is true, whereas the visual test performs well.

2.6.6 Do people tend to pick the lowest p-value?

One assumption made in order to evaluate the effect of lineup size in the calculations of visual p-value and signal strength was that subjects would tend to pick the plot in the lineup that had the strongest signal. In experiments 1 and 2, this corresponds to the plot with the smallest p-value. We examine the data collected from the first two experiments, to see if this assumption is, indeed, reasonable.

Figure 2.10 gives an overview of all selections in all lineups of experiments 1 and 2. Each panel of the figure corresponds to a single lineup. Each 'pin' – a short line topped by a dot – corresponds to one plot in the lineup. The x-location of the pin shows the plot's p-value on a log scale, its height is given by the number of observer choosing this plot. Columns are ordered

according to effect size as defined in section 6.2; rows show replicates for the same combination of parameters.

Red indicates the plot with the lowest p-value in the lineup. Blue indicates the plot of the actual data when it is different from that with the lowest p-value. In both experiments people tended to select the plot with the lowest p-value. The results are clearer for experiment 2, that used a continuous covariate. But even when subjects did not pick the plot with the lowest p-value they tended to oscillate their choices between the several low p-value plots. So for most subjects, the assumption that they pick the plot with the smallest p-value would appear to be reasonable, and the actual power of the visual test should be close to the expected power.

There are some noticeable exceptions to this rule. In experiment 1, when $\beta = 0, n = 100, \sigma = 5, rep = 1$ people overwhelmingly chose a plot with much larger p-value, similarly, for parameters $\beta = 5, n = 300, \sigma = 12, rep = 3$, people tended to pick the plot with the second smallest p-value. For several of these exceptions, along with several easy lineups, a follow up experiment was conducted using an eye-tracker to examine which patterns or features participants are cueing on in making their choices (Zhao et al., 2012).

2.6.7 How much do null plots affect the choice?

Visual inference falls into the same framework as randomization tests, where the statistics from the data are compared with those from null data. Unlike randomization tests visual inference is constrained to make the comparison with just a few draws $(m - 1)$ from the null distribution. How this small set of null plots influences the subjects' choice is important for understanding the reliability of visual inference. If the actual data plot is very different from all of the null plots, then the null plots should not have much influence on the choice. Measuring the difference, generally, between plots is almost impossible. However, in this controlled setting we can use p-values of the test statistic calculated on the data used in each plot as a proxy for similarity of structure between the plots. If there is a null plot with a small p-value, or one close to that of the actual data plot, we would expect that subjects have a harder time detecting the actual data plot.

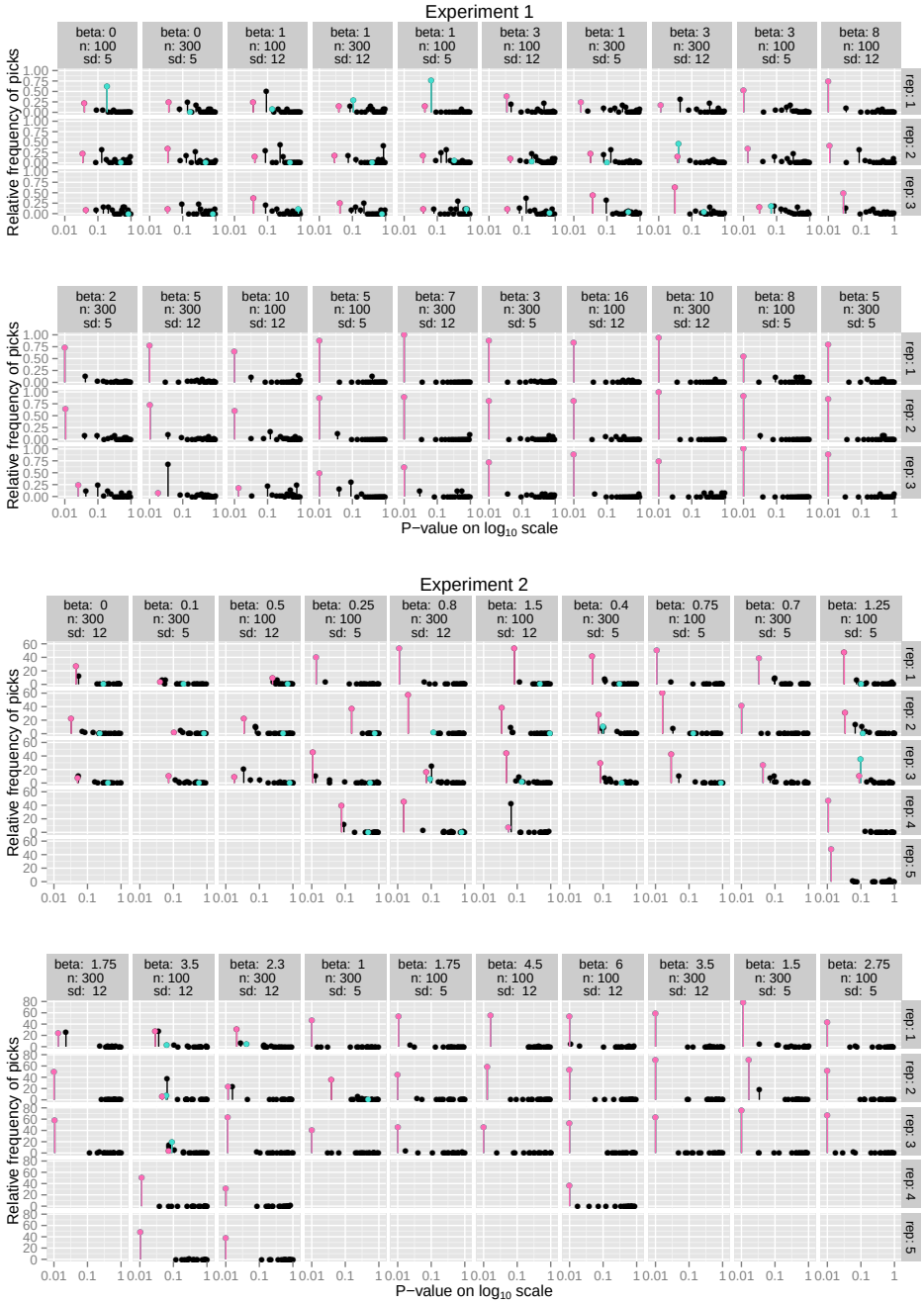


Figure 2.10 Relative frequency of plot picks compared to other plots in the lineup plotted against the p-value (on $\log_{10}$ scale) of each plot for all individual lineups of both experiment 1 and 2. Red indicates the plot with the lowest p-value, and blue indicates the actual data plot, when it is different from that with the lowest p-value. Columns are ordered according to effect size, with rows showing replicates of the same parameter combination on top of each other. Empty cells indicate combination of parameters that were not tested. Highest counts tend to be the plot in the lineup having the lowest p-value, more so for experiment 2 than 1.

2.6.8 Type III error

A little known error amongst statisticians is what was coined as Type III error in Mosteller (1948). Type III errors are defined as the probability of correctly rejecting the null hypothesis but for the wrong reason. Experiment 3 is prone to this type of error. Participants were asked to identify the plot with the largest absolute slope. But the actual data plot featured a cluster of points, the contamination that made the conventional test fail to see any trend. For the human eye this cluster of points is as visible as the association between the remaining points, enabling the observer to identify the actual data plot by looking for the cluster instead of the slope. This would be considered a Type III error because it leads to a correct rejection of the null hypothesis, but is not related to the value of the slope parameter.

For visual inference, making a Type III, is not actually a problem. It is only a possibility in this experiment because we are working with known structure. In the real setting, we are excited to see observers detecting the actual data plot, and curious about how they detect it, with all possible reasons encapsulated in the alternative hypothesis. However, this highlights the importance of getting qualitative reasoning from observers for their choices.

2.7 Conclusions

This paper has demonstrated that statistical graphics can be used in statistical inference and validates the lineup protocol proposed by Buja et al. (2009). Specific terminology was defined, and methods for obtaining the p-value and estimating the power of visual tests were introduced. In order to calculate the theoretical power, it was assumed that observers will select the plot having the strongest signal in the lineup, and the experimental data suggests that for most observers this assumption holds. Results from visual inference in the controlled setting of the simulation study are comparable to those obtained by conventional inference. Visual inference is intended to provide valid tests where no conventional test exists, and our experiments in a controlled scenario suggest that it will perform as expected in the intended applications. The power of a visual test increases with the number of observers, which interestingly, leads to a

result that the theoretical power of visual test can be better than that of conventional tests.

The lineup protocol operates similarly to statistical tests that have broad alternative hypotheses. If the null hypothesis is rejected, generally we can say that "there is something there" but not specifically what it is in the data that triggers the rejection. Follow-up questions on the reasons provide qualitative insight. In conventional testing, multiple comparisons are often done to refine and understand the test results, and perhaps some similar approaches might be developed for visual inference.

The performance of subjects was quite varied, but consistent. No restrictions were placed on Turk workers, in terms of abilities. There were clearly some subjects who performed very badly, but it was very interesting to see that there were some super-observers, people who detected the actual data plot at a rate better than that of the power of the best conventional test. It would be interesting to see how well trained subjects might perform. Prior to the Turk experiments, we conducted pilot studies using local graphics experts and obtained good results, indicating that training in data visualization might be helpful for visual inference. Future work might explore this.

Visual inference has been successfully used in two practical applications: to evaluate the power of competing graphical designs (Hofmann et al., 2012), and to detecting signal presence in large p, small n data (Roy Chowdhury et al., 2011). It is hoped that the lineup protocol will prove to be valuable in data mining applications, and exploratory analyses, where there are no existing gauges of statistical significance.

Supplementary Material: Proof of Lemma 2.3.1, details of data collection and cleaning, longer discussion of effect of null plots and Type III error.

CHAPTER 3. HUMAN FACTORS INFLUENCING VISUAL STATISTICAL INFERENCE

A paper to be submitted to *Sociological Methodology*

Mahbubul Majumder, Heike Hofmann, Dianne Cook

Abstract

Visual statistical inference is a way to determine significance of patterns found while exploring data. It is dependent on the evaluation of a lineup, of a data plot among a sample of null plots, by human observers. Each individual is different in their cognitive psychology and judiciousness, which can affect the visual inference. The usual way to estimate the effectiveness of a statistical test is its power. The estimate of power of a lineup can be controlled by combining evaluations from multiple observers. Factors that may also affect the power of visual inference are the observers' demographics, visual skills, and experience, the sample of null plots taken from the null distribution, the position of the data plot in the lineup, and the signal strength in the data. This paper examines these factors. Results from multiple visual inference studies using Amazon's Mechanical Turk are examined to provide an assessment of these. The experiments suggest that individual skills vary substantially, but demographics do not have a huge effect on performance. There is evidence that a learning effect exists but only in that observers get faster with repeated evaluations, but not more often correct. The placement of data plot in the lineup does not affect the inference.

Keywords: statistical graphics, non-parametric test, cognitive phycology, data visualization, exploratory data analysis, data mining, visual analytics.

3.1 Introduction

The lineup protocol introduced in Buja et al. (2009) can be used to test the significance of findings during the exploratory data analysis. The methodology is a part of what is called visual statistical inference. These concepts have been developed further by Majumder et al. (2013b) who refined the terminology and validated the lineup protocol with a head to head comparison with conventional inference. One of the major contributions of Majumder et al. (2013b) is to define the power of the visual test and provide methods to obtain the power for a particular lineup. It was observed that the power can be as good or better than that of a conventional test in some scenarios.

In visual inference, the test statistic is a plot of the observed data. To create a lineup, this plot, called the actual data plot, is placed in a layout of null plots. The null plots are generated from the model specified by a null hypothesis, essentially describing what the plot might look like if the data had no structure. An observer is asked to evaluate the lineup. If the actual data plot is detected by the observer, the null hypothesis is rejected. This means that the structure in the actual data plot has significant structure, a pattern that is not simply due to randomness. Combining the choices of multiple observers provides more stability in the estimation of significance.

Figure 3.2 displays a lineup of 20 plots where one of the plots is observed data, while the remaining 19 plots are rendered from data generated under a null model. Which one of the 20 plots is the most different from the others? When asked this question, 12 of 72 observers picked the same plot (with number equal to the result of $3^2 + 4$), with reasons given being 'asymmetry' (36%), 'trend' (26%) or 'outliers' (16%). The corresponding p-value is 0.00023, indicating sufficient evidence to reject the null hypothesis.

What does this mean, though? For that, we need to know the context of the data and we need to have more information about the generation of the null plots. For that, we need to know the context of the data and we need to have more information about the generation of the null plots. This example investigates the results from the 2012 US presidential election in comparison to the poll results just prior to the election. (Although this example is more

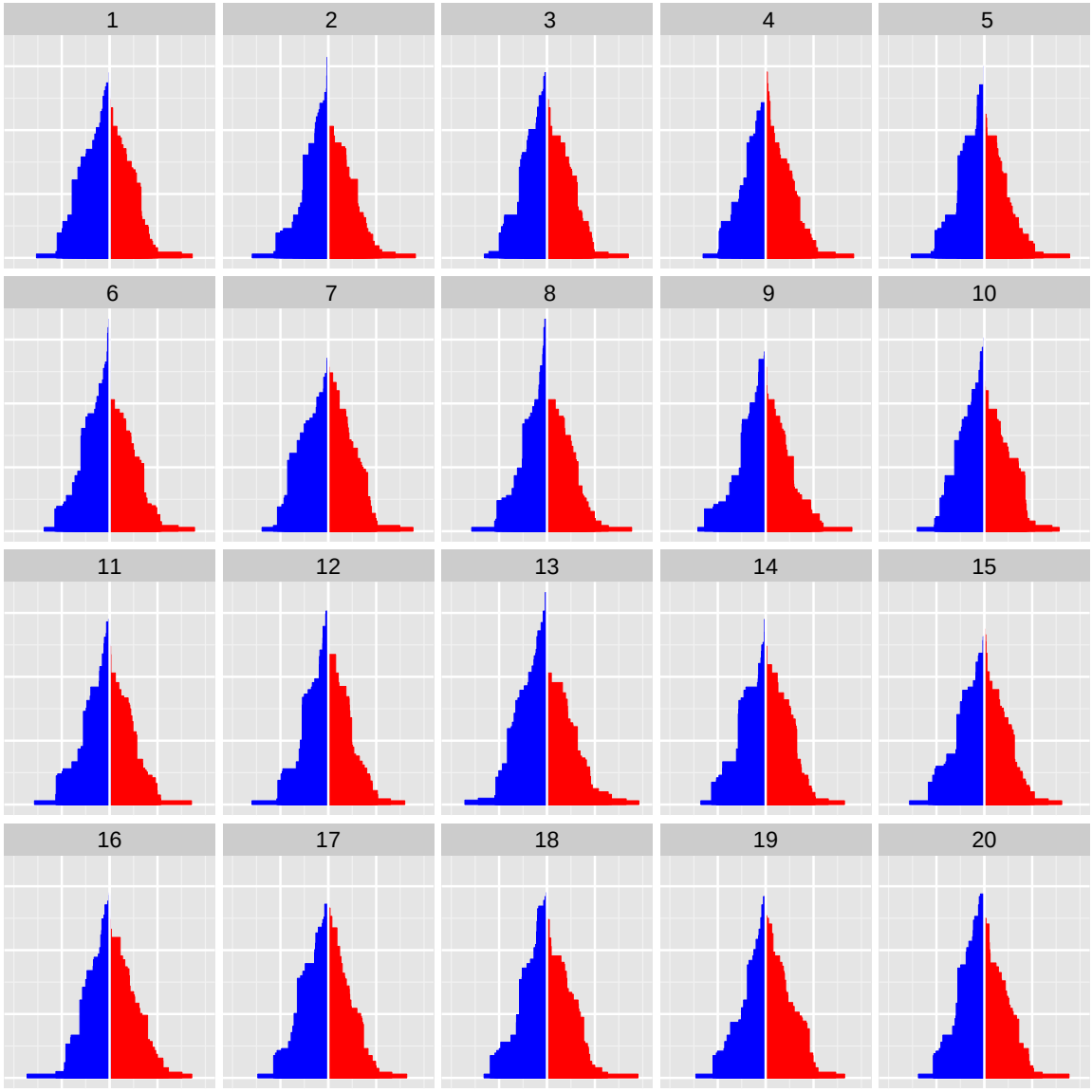


Figure 3.1 Which one of the plots is the most different from the others?

simplistic than most of the tests conducted to date, it will serve the purpose of illustrating the lineup protocol.) The data is looking at the difference in poll results between the two (major) presidential candidates, Obama and Romney, for all states. Each panel in Figure 3.1 shows an 'electoral building' where each state in the union is represented by a rectangle. This difference is plotted horizontally, and the height of each box corresponds to the state's electoral votes. Color indicates party affiliation. The null hypothesis is "that the election results were consistent with the polls". The polling results provides the null model from which data is simulated. Because each poll has a margin of error, this is used to simulate different scenarios that might have resulted on election day, if the polls were on target. A null data set is generated as a set of draws from normal distribution, with mean equal to the difference in poll percentage of the latest state poll results, and standard deviation equal to 2.5, approximating a margin of error of 5%. These samples are plotted as electoral buildings, and the plot made with the results from the election is placed randomly among them in a lineup of size 5×4. If the null hypothesis is true the actual data plot should look just like any of the other plots, and not be identified by an observer. Figure 3.2 shows a plot of the electoral building with added context information and labels.

A lineup can be evaluated by a single person or multiple observers. A binomial distribution is used to calculate the p-value based on the number of times observers identify the actual data plot, which provides the information needed to make a decision on rejecting or failing to reject the null hypothesis. Observers should not be aware of the data that constitutes a lineup, and should not have seen the actual data plot before seeing the lineup. This is the reason that in the election example, above, the scenario was explained after the lineup question, in the text.

The question that is asked of the observer should be as general as possible, effectively asking the observer to pick the plot that is different, and allowing them to provide their reasons for seeing their pick as different. In some of the studies, the ones described in Majumder et al. (2013b) very specific questions were asked, because the experiments were being conducted to compare results from the lineup protocol with those of conventional tests. In those experiments, structure in the data was strictly controlled in the simulation process, which allowed for specific questions to be asked. In the election example, observers were asked "which plot is the most

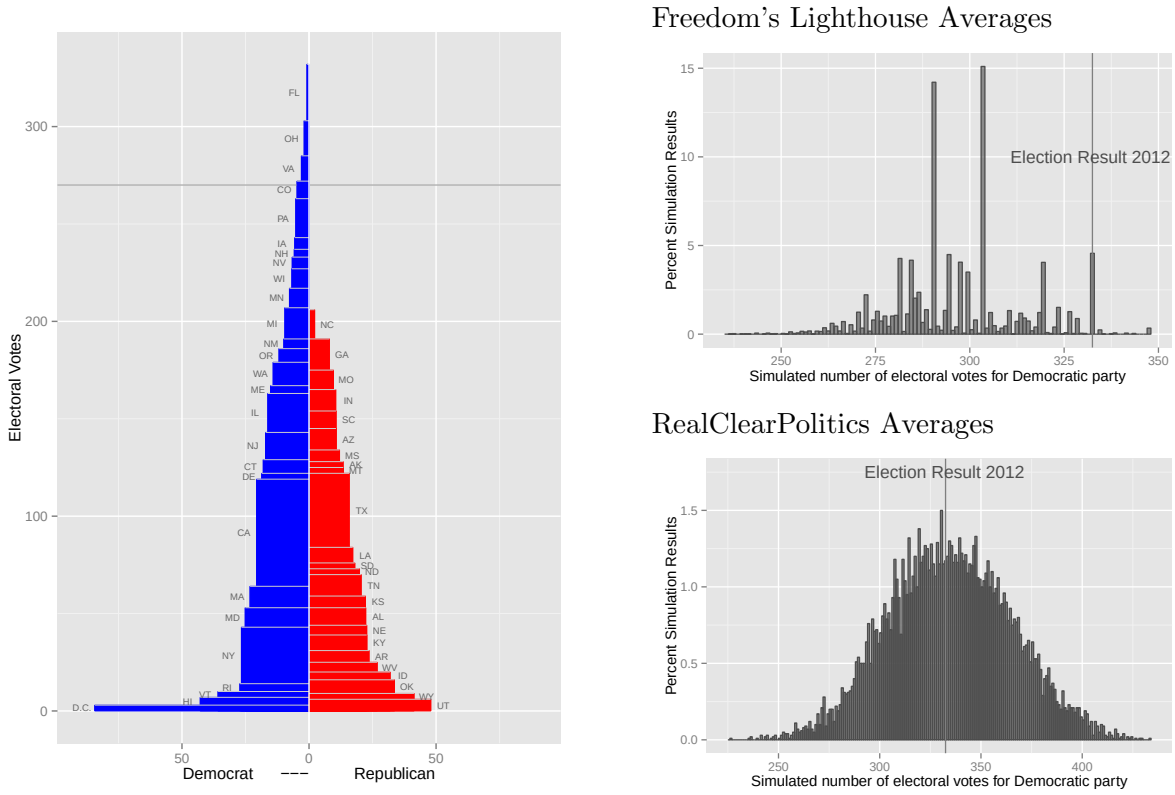


Figure 3.2 Electoral building plot of the results of the 2012 U.S. Presidential Election (left). On the right two histograms of 10,000 simulations each based on polling averages from two different sources. For the histogram on the top, the p-value of observing results as extreme as the 2012 U.S. election results based on the bootstrap is 0.0533 (with Bootstrap standard error of 0.002), making the election results almost significantly different from the polls. There is no indication of any inconsistency between polls and election results based on the bootstrap simulation below. The lineups are based on the top source.

different?". The type of plot, showing two (modified) stacked bar charts in different colors should suggest to the observer that the interesting feature is most difference between the two heights. Most observers got this, but it is possible that some observers might pick plot 4, where the red tower is slightly above the blue as the most different because it is the only plot with this feature. So a better question may have been "Which plot shows the biggest height difference between the two towers?" except that this tailors the inference to a specific feature which does not match the null hypothesis of interest.

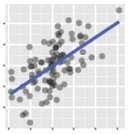
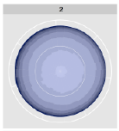
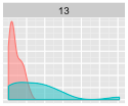
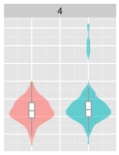
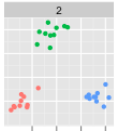
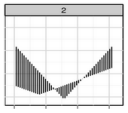
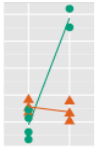
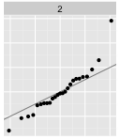
For any evaluation the observer may or may not identify the actual data plot. Under the null hypothesis, the actual data plot should look similar to the null plots making it harder to detect. It is not expected that an observer would be able to detect the data plot in this scenario. But since there are limited number of plots in a lineup, which is 20 in the election example, there is a 1/20 chance that the observer would pick the actual plot. This proportion is associated with the Type I error of the test. On the other hand, if null hypothesis is not true, the observed plot should look different from the null plots, making it easier to be detected. This is the definition of the power of the test. When multiple observers evaluate a lineup, the proportion of correct response can be used to estimate the power. The ability of individual observers can vary, and examining this is the purpose of this paper.

There have been ten experiments conducted using Amazon's Mechanical Turk (Amazon, 2010) that are being used to evaluate the effects of observers' demographic factors on the inference. Table 3.1 summarizes these experiments. Each of these experiments collected demographic details of the subjects. Experiments 5, 6 and 7 are used to study the learning trend of the observer. The design of experiment 9 incorporated components that allows position of the actual data plot in the lineup, and the sample of nulls, to be evaluated. Section 3.2 discusses human factors that may affect the performance of the observer. Section 3.3 describes the methods used to assess the effects, and Section 3.4 describes the results.

3.2 Factors Affecting Observer Performance

Based on human evaluation on a lineup a decision is made in visual statistical inference. While the visual test statistic is defined to have a direct influence on the observer to pick a plot

Table 3.1 Visual test statistics used in 10 different simulation experiments. The observers are asked different questions to answer while evaluating a lineup

Serial	Experiment	Test Statistic	Lineup question
1	Box plot		Which set of box plots shows biggest vertical difference between group A and B?
2	Scatter plot		Of the scatter plots below which one shows data that has steepest slope?
3	Contaminated plot		Of the scatter plots below which one shows data that has steepest slope?
4	Polar vs Cartesian		Which plot is different?
5	Hist vs density		In which plot is the blue group furthest to the right?
6	Violin vs boxplot		In which plot does the blue group look the most different from the red group?
7	Group separation		Which of these plots has the most separation between the coloured groups?
8	Sine Illusion		In what picture is the size of the curve most consistent?
9	Gene expression		In which of these plots is the green line the steepest, and the spread of the green points relatively small?
10	Test normality		Which of these plots is most different from the others?

based on the signal in the data, there are other human factors that may affect the observer performance. It is important to study the effect of those factors and examine the extent of their influence. This section presents a brief description of the factors that may affect the power of visual inference.

3.2.1 Signal in the Data

The visual test statistic is chosen so that it displays a specific pattern in case the null hypothesis is not true. Thus the most important factor that help an observer to correctly evaluate a lineup is the presence of any detectable signal in the data. On the other hand, if the null hypothesis is not true, visual test statistic should not display any distinguishable pattern. In fact, some of the null plots may appear to be the most different plot in a lineup influencing the observer to chose a plot different than actual data plot. This is an elegant feature of the lineup. It force the observer to chose a wrong plot when the null hypothesis should not be rejected.

3.2.2 Choice of Visual Test Statistic

The visual test statistic should be highly associated with the hypothesis under investigation. To achieve this purpose it is very important to decide which plot type and plot features should be adopted. In a linear regression setting, the visual test statistics are presented in (Majumder et al., 2013b) for some common hypothesis testing. It is also observed that a scatterplot may do a better job than a box plot when using as a visual test statistic for regression parameters. Some of the effective features of visual test statistics are discussed in (Hofmann et al., 2012) including plot type, color and shape of the plots. Roy Chowdhury et al. (2012) presents some *distance measure* to determine how a plot may be different from each other.

Table 3.1 shows the visual test statistics used in this study. For both experiments 2 and 3 a scatterplot is used but a regression line fitted through the points is overlaid for experiment 2. It is easier to spot the slope of the line compared to just the scatterplot itself. But people are better in noticing the unusual pattern in the data which is some contamination purposefully added for experiment 3. For experiment 3 the performance of the observers was much better

compared to experiment 2 (Majumder et al., 2013b).

3.2.3 Question that Human Observer Answers

The researcher knows about the underlying hypothesis but the observer does not necessarily know the underlying details of the lineup. So, the researcher needs to ask a question to the observer to answer while evaluating the lineup. This question should provide the observer a little clue so that the answer reflects the hypothesized patterns in the actual plot. For example Table 3.1 shows the questions asked for the simulation experiments considered for this study. Notice that for case 1 if the observer can identify the actual plot in the lineup that should indicate that the plot chosen has the most vertical difference between groups A and B which is exactly what the researchers intend to examine. Similarly for case 2, a correct evaluation would indicate that the slope is different than the slope that may show up just from randomness.

These questions are very crucial for the power of visual test. They help observer think in a meaningful direction. Notice that there may be unnecessary patterns in the actual data plot which may not necessarily indicate the existence of the significant signal in the plot. These questions help observer not to be misguided by those patterns. To follow up further on this Majumder et al. (2013b) also collected reasons of choice made by the observers and noticed that Type-III error may occur where the hypothesis may be correctly rejected but for a wrong reason.

3.2.4 Demographics of the Observer

Some of the demographics of the observer such as age, gender, education level and geographical location may have effect on how an observer would examine the lineup. This may produce some variability in the performance of the observer. For example, a high school student may not respond same as a well educated person while evaluating a lineup. There may be variations in different age groups as well. Investigation on subjects with a well variety of age groups is necessary to study this variability.

To meet the Institutional Review Board (IRB) requirement, it is necessary to exclude any subject less than 18 years of old to participate human subject experiment. Thus we intend to

only focus on the performance of observer with age 18 years or older. We don't expect many older people to participate our study and we are interested of age from 18 to 65 years.

Education level has some relation with age. Younger observer may not have higher degree. Thus some of the variability in performance for different education level may be confounded with age. On the other hand, it could be possible to have undergrad degree with higher age. Specially for any geographical location where not many certain gender group have higher education, this may occur. Thus the effect of all these demographical factors may be confounded at certain level.

We don't expect much difference in performances between male and female observers. Gender may have some influence through the education level and geographical locations since some location may have small number of educated female population. We intend to include similar number of male and female subjects in our study so that this can be examined.

3.2.5 Learning Trend of the Observer

When an observer evaluates a lineup he or she may learn something from the experiences of the earlier observations. If the same observer is shown another lineup the learning from previous evaluation may help. The more evaluations an observer makes the more skillful the observer may become. This learning trend may or may not be significant overall but it may influence the performance of the observer in some way.

Learning may occur in two different ways. An observer can become more trained on the lineup structure and the pattern in the plots shown. This can be learning in proportion of correct evaluation. Or, he or she may become efficient in responding faster in the later evaluations. This can be observed as the time taken for evaluating a lineup. This paper investigates both of these learning trends with multiple simulation experiments.

When an observer evaluate a lineup for the first time, he or she has to read through many instructions and become familiar with the environment of the experiment. We expect that this will require much more time compared to the rest of the evaluations. Time taken for the second or successive attempt should be much lower than the time taken for the first lineup. We attribute this to the adaptation to the experiment environment, not a learning of evaluating a

lineup.

3.2.6 Location of Actual Plot in the Lineup

The actual data plot is placed in a random spot in a lineup. While evaluating the lineup some people may start looking from some specific part of the lineup. With the help of eye-tracking equipments Zhao et al. (2012) tracked the observers' eyes to see how they went through the plots in a lineup to come to their answers. The results suggest that people have particular methods of reading lineups. Some people read lineups from left to right direction while some read from upward to downward. Some people start looking from the center of the lineup while others start from the top left corner. In the earlier phase of the exercise, the observer tend to scan the plots and start comparing plots to make a final decision. Beside right-left or up-down directions observers show some diagonal movement too.

Given a specific pattern of eye movement of the observer in examining the lineup, the location of actual plot in the lineup may have some effect. Those who start exploring from the left top corner may get first glance of the actual plot if it is on that location. This should give the observer more time to compare it with the rest of the null plots. Those who start looking from the center may scan towards right or left direction and thus don't have the chance to scan the null plots continuously as they could if they would start from a corner. Thus they should scan the center plot over and over again while scanning the left or right side plots. If the data plot is in the center location it may have multiple chance to be examined and hence be identified.

It may be possible that some part of the lineup is less explored or even never scanned by the observer if he or she feels that the actual plot is found before even seriously scanning the whole lineup. In those scenarios the location where the observer first start scanning is important. Placing the actual data plot in that location may yield different results.

3.2.7 Selection of Null Plots

A lineup becomes difficult to evaluate if one of the null plots appears to be very similar to the actual plot. When null hypothesis is true it is a common scenario. With alternative

hypothesis being true, it may happen if we compare actual plot with many null plots. But we have only specific number of null plots in a lineup. So, a different set of null plots may yield some variation in the difficulty of the lineup with the same actual plot. This may affect the performance of the observer while evaluating a lineup.

3.2.8 Individual Performance of the Observer

Each person is different from others in some way. For example, in a controlled experimental study Zhao et al. (2012) noticed that some people spent a lot of time to decide no matter whether the lineup is difficult or easy while some simply glanced at lineups to make a decision. This influences the response of the observer. Also subject specific variation in the power of visual test is observed in Majumder et al. (2013b).

3.3 Experimental Designs and Methods

While evaluating a lineup, the performance of the observer may depend on the factors described in Section 3.2. Some of these factors such as signal in the data and individual abilities were studied in Majumder et al. (2013b). In this paper we intend to study effect of the factors such as human demographics, learning trend, actual plot locations and selection of null plots. This section presents the simulation experiment setup, data collection methods and data analysis plans with models to asses the influence of those factors on visual statistical inference.

3.3.1 Experiment Setup

In addition to obtain evaluation responses of the lineups, all the 10 experiments shown in Table 3.1 were designed to collect the following demographic information of the subjects participating the experiments:

1. Age group
2. Gender
3. Academic education level

4. Geographical location

Instead of collecting exact age, 9 levels of age were collected. They are 18-25, 26-30, 31-35, 36-40, 41-45, 46-50, 51-55, 56-60, above 60. There were five levels of academic information. They are (1) High school or less , (2) Some under grad course, (3) Under graduate degree, (4) Some graduate courses, (5) Graduate degree. Geographical locations were collected using the true public ip addresses of the participants' computer. This provides latitude and longitude of the ip address as well as the city and country information.

3.3.1.1 Learning Trend

Learning trend of a subject can be observed in terms of performance over successive feedbacks received when multiple lineups are shown for evaluation. Experiments 5, 6 and 7 were used for this. Each subject was shown a total of 10 lineups randomly selected from a pool of lineups. The lineups are not necessarily with the same difficulty levels. But the order of lineups were randomized. The responses of the lineups were recorded by attempt 1 through 10. Attempt 1 means that the response is for the first lineup the observer evaluates and attempt 10 refers to the response for the 10th lineup. The goal is to estimate whether performance of the observer improves from attempt 1 to attempt 10.

3.3.1.2 Design for Location Effect

Experiment 9 shown in Table 3.1 is designed to to assess the location effect of a data plot in a lineup. It is set up based on the gene expression data (Atwood et al., 2013) with two groups. One is the Genotype and the other the Empty Vector (EV). For each of the groups, gene expression data were collected in presence or absence of iron sufficiency. Two factors were of primary interest. One is the Genotype main effect and the other is the interaction effect between Genotype and iron sufficiency.

For Genotype main effect side by side dot plots of two groups are used as the visual test statistics. For interaction effect the same plot type is used except that the means of iron sufficiency or insufficiency were connected by lines colored by the groups. One of the interaction

test statistics is shown in Table 3.1 for experiment 9. For no interaction, the visual test statistic shows two parallel lines representing two groups and for significant interaction the lines are not parallel. Null plots were generated by randomizing the group structure of the data. The actual plot was randomly placed in five different locations in a lineup of size 20. The locations are 2,9,12,16,20 for Interaction effect and 1,8,12,17,20 for Genotype effect. For a lineup with specific data plot location, five different sets of null plots were used to produce 5 lineups for each location. In total we have 25 lineups for Interaction effect and 25 lineups for Genotype effect.

Each observer saw three lineups, first an Interaction lineup then a Genotype lineup. Finally a test lineup was shown to screen out unusual evaluation. The subjects did not know which one was the test lineup but they were informed before accepting the task that there would be one. The test lineup was very easy and everyone should detect the data plot. The MTurk workers were paid based on whether they could correctly evaluate the test lineup. But the data on the test plot are excluded from the analysis.

3.3.2 Data Collection Methods

Human subjects were recruited to evaluate the experimental lineups through Amazon Mechanical Turk (Amazon, 2010) or MTurk Web site. It is an online work place where people from around the world can perform some tasks and get paid. Usually tasks are very simple and no specialized training is required to do them. Being a human is the main requirement. Tasks are designed for anyone to do but some tasks may require some skills depending on the recruiters' need. Each task is usually planned to complete in a short time. Humans are still better than computers in performing these types of tasks. The the amount of money paid for each task is very small as well.

We designed and developed a web application which enables to display the lineups to the observers as per experimental need. The MTurk workers were redirected to the web site to evaluate lineups. The data were collected, stored automatically into a local database server. Demographic informations such as age group, gender and education levels were also collected. The time taken for each evaluation is computed based on the time the plot was shown and

the time the feedback was received. It is measured in seconds. The location of the observer is determined by the ip address of the observer.

3.3.3 Model to Estimate Demographic Factor Effect

The feedback provided by each observer on a lineup is a binary response variable. Suppose Y_{ij} denotes the response from observer i on a lineup j. $Y_{ij} = 1$ if the response is correct otherwise $Y_{ij} = 0$. Let $\pi_{ij} = E(Y_{ij})$ be the probability that observer i correctly picks the data panel from lineup j. We model this in a generalized mixed effects model of the form

$$g(\pi_{ij}) = \mu + \alpha_{k(i)} + \gamma_{l(i)} + \tau_{m(i)} + \kappa_{s(i)} + \ell_j, \tag{3.1}$$

where μ is an overall average probability for picking out the data plot from a lineup. α, γ, τ and κ are the effects of age group $k(i)$, gender category $l(i)$, education level $m(i)$ and country name $s(i)$ respectively for observer i. ℓ_j is a random intercept predicting lineup difficulty level and we assume $\ell_j \sim N(0, \sigma_\ell^2)$. $g(.)$ denotes the *logit* link function $g(\pi) = \log(\pi) - \log(1-\pi); 0 \leq \pi \leq 1$.

The effect of demographic factors can be observed as time taken for each evaluation by an observer. Suppose Z_{ij} denotes the logarithm of time taken for an observer i to evaluate a lineup j. Let $\mu_{ij} = E(Z_{ij})$ be the average of log(time taken) by observer i to pick the data panel from lineup j. We model this in a mixed effects model of the form

$$Z_{ij} = \mu + \alpha_{k(i)} + \gamma_{l(i)} + \tau_{m(i)} + \kappa_{s(i)} + \ell_j + \epsilon_{ij}, \tag{3.2}$$

where μ represents overall average of log time taken by an observer to evaluate a lineup. α, γ, τ and κ are as described in Model (3.1). ℓ_j is a lineup-specific random effect for the time needed to evaluate a lineup; $\ell_j \sim N(0, \sigma_\ell^2)$ and the overall error $\epsilon_{ijk} \sim N(0, \sigma^2)$.

3.3.4 Model to Estimate Learning Trend

Multiple lineups were sequentially shown to each observer for evaluation. Suppose Y_{ijk} denotes the response from observer i on a lineup j in his/her kth attempt. $Y_{ijk} = 1$ if the response is correct otherwise $Y_{ijk} = 0$. Let $\pi_{ijk} = E(Y_{ijk})$ be the probability that observer i correctly picks the data panel from lineup j in his/her kth attempt. We model this in a

generalized mixed effects model of the form

$$g(\pi_{ijk}) = \mu + \alpha_k + u_i + a_i k + \ell_j, \tag{3.3}$$

where μ is an overall average probability for picking out the data plot from a lineup, α_k is the effect of the kth attempt on the probability, with $\alpha_1 = 0$ and $k = 1, ..., K$. u_i and a_i are observer specific random effects, $i = 1, ..., I$. u_i is a random intercept, describing a basic subject-specific ability. We assume $u_i \sim N(0, \sigma_u^2)$. a_i is a random slope capturing an individual's specific learning effect over the course of K attempts. We assume $a_i \sim N(0, \sigma_a^2)$. For ℓ_j we again assume a normal distribution, $N(0, \sigma_\ell^2)$. ℓ_j is a random intercept predicting lineup difficulty level. $g(.)$ denotes the *logit* link function $g(\pi) = \log(\pi) - \log(1 - \pi); 0 \leq \pi \leq 1$.

The inverse link function, $g^{-1}(.)$, from Equation 3.3 leads to the estimate of the subject and the lineup specific probability of successful evaluation in kth attempt by a single observer as

$$\hat{p}_{ijk} = g^{-1}(\hat{\mu} + \hat{\alpha}_k + \hat{u}_i + \hat{a}_i k + \hat{\ell}_j). \tag{3.4}$$

The learning of each observer over a course of K evaluations may be observed as the improvement of the time taken to evaluate a lineup in the later attempts. Suppose Z_{ijk} denotes the logarithm of time taken for an observer i to evaluate a lineup j in his/her kth attempt. Let $\mu_{ijk} = E(Z_{ijk})$ be the average of log(time taken) by observer i to pick the data panel from lineup j in his/her kth attempt. We model this in a mixed effects model of the form

$$Z_{ijk} = \mu + \alpha_1 + \alpha k + u_i + a_i k + \ell_j + \epsilon_{ijk}, \tag{3.5}$$

where μ represents overall average of log time taken by an observer to evaluate a lineup. α is the average change in log time taken for each additional attempt. α_1 is an offset in log time taken for the first attempt. All other effects are random effects: as before, u_i is a subject-specific intercept representing individual speed of an observer with $u_i \sim N(0, \sigma_u^2)$. a_i is a subject-specific slope representing the deviation of the speed-up (or -down) by attempt k. We assume $a_i \sim N(0, \sigma_a^2)$. ℓ_j is a lineup-specific random effect for the time needed to evaluate a lineup; $\ell_j \sim N(0, \sigma_\ell^2)$ and the overall error $\epsilon_{ijk} \sim N(0, \sigma^2)$.

Equation 3.5 leads to the estimate of the subject and the lineup specific time taken for an evaluation in kth attempt by a single observer as

$$\hat{\mu}_{ijk} = \hat{\mu} + \hat{\alpha_1} + \hat{\alpha}k + \hat{u}_i + \hat{a}_i k + \hat{\ell}_j. \tag{3.6}$$

To fit all these mixed effect models the function $lmer()$ is used from R package $lme4$ by Bates et al. (2011). To obtain the p-values of fixed effect parameters estimates, normal approximation is used for Z scores computed as the ratio of estimates to the estimated standard errors.

3.3.5 Model to Estimate Location Effect

As per the design of experiment for location effect, the actual data plot is same for each of the null sets of plot. Thus the response data of this experiment constitute a multivariate response. To examine if the difference in proportion correct among the locations is statistically significant we fit a one way multi variate analysis of variance (MANOVA) model to the data.

Suppose $\mathbf{Y} = (Y_1, Y_2, ..., Y_p)^\top$ be a vector of random variable with dimension p, the total number of null sets. let $\mathbf{Y}_{ij}$ represents jth vector response for ith location with $i = 1, 2, ..., I$. We fit the following MANOVA model

$$\mathbf{Y}_{ij} = \mu_i + \epsilon_{ij} \tag{3.7}$$

where $\mu_i = (\mu_{1i}, \mu_{2i}, ..., \mu_{pi})^\top$ is the mean vector for location i and $Var(\epsilon_{ij}) = \Sigma$.

3.4 Results

3.4.1 Overview of the Data

A total of 2321 participants provided feedback data on the lineups in ten different experimental studies. Figure 3.3 displays the locations of participants around the world. Most of the participants were from the United States and India. There were 76 other different countries from where we received feedbacks. This provides a diverse pool of participants. The diversity in not only in geographical locations of the participants but also in their gender, age groups and education levels as we see in Table 3.2. It is interesting that there were large number of female participants even though there were lot of people from developing countries.

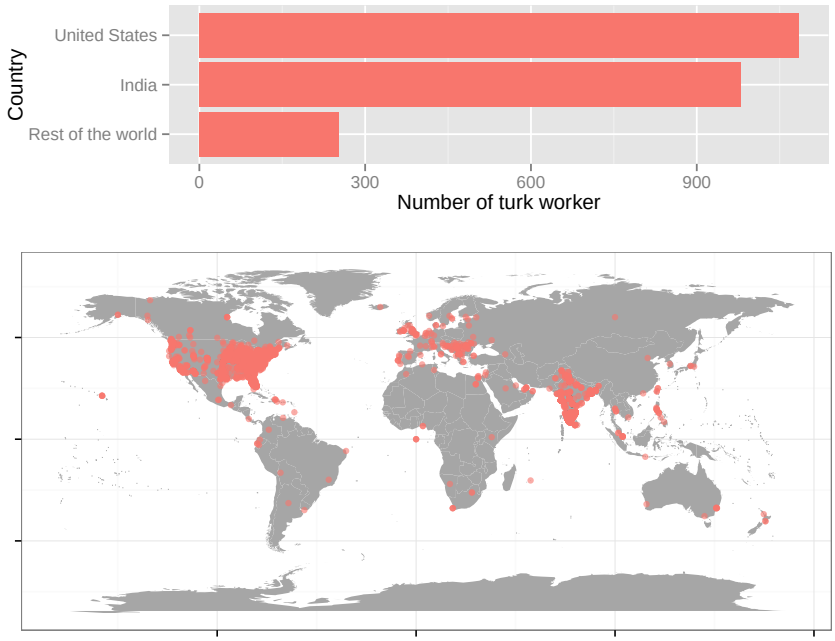


Figure 3.3 Location of the Amazon Mechanical Turk workers participating our study. Most of the people are coming from India and United States even though there are people from around the world.

Besides United States and India, countries such as Canada, Romania, United Kingdom and Macedonia have more than 10 participants each. The rest of the 70 countries have less than 10 participants. The distribution of participants remains almost similar in all ten experiments with some small variations (Figure B.2). That may be due to the time of the experiment when it first started. For some experiments India got more participants than United States and for some experiments the numbers just got reversed (Figure B.3). It is also observed in some experiments that number of participants are similar no matter whether the experiment is run on day time or nigh time.

The largest number of participants are from age group 18 to 25 which is the youngest age group in the study. The majority of the people have ages between 18 to 35. Interestingly there are many participants from older age groups as well. Specially for united states almost all the age groups show uniform participations beyond age 30 (Figure B.4). Notice that fewer people participated from India beyond age 40 compared to united states. Total number of participations from india and United States are almost same with United States having 107

Table 3.2 Demographic information of the subjects participated the MTurk experiments. Average time taken for evaluating a lineup is shown in seconds.

		Participants		Average	
Factor	levels	Total	%Factor	time	response
Gender	Male	1348	57.63	48.51	13493
	Female	991	42.37	43.75	10564
Education	High school or less	193	8.24	37.21	2241
	Some under graduate courses	418	17.85	42.84	4070
	Under graduate degree	584	24.93	44.29	5775
	Some graduate courses	245	10.46	43.43	2460
	Graduate degree	902	38.51	52.18	9511
Age	18-25	740	31.61	42.97	7311
	26-30	547	23.36	46.27	5585
	31-35	376	16.06	44.27	3923
	36-40	257	10.98	55.03	2714
	41-45	141	6.02	43.90	1519
	46-50	95	4.05	49.29	1003
	51-55	83	3.54	48.67	867
	56-60	64	2.73	59.73	678
	above 60	38	1.62	48.67	457
Country	United States	1087	46.83	39.64	10769
	India	980	42.22	52.63	10227
	Rest of the world	254	10.94	46.86	2819

more participants. Unlike USA Indian participants are mostly young people.

In terms of participant's academic background, the largest group is graduate degree. A total of 902 participants have a graduate degree which is about 38.51%. They are mostly from India as any university degree is considered a graduate degree unlike north America where graduate degree means Masters level education. Most of the USA participants are with an Undergraduate degree or at least have some undergraduate courses (Figure B.4).

The distribution of male and female participants are similar among all age groups except age 18-25 in India where fewer female participants were observed (Figure B.5). The distribution of education levels are also different across the countries for this age group. Most of the participants from India are below age 40 while in United States the distribution of participants are similar beyond age 40. There were not many participants beyond age 50 and hence these age groups are merged to form one age group called above 50. The exploratory analysis and the models fitted in the following sections use this new age group instead of actual age groups beyond age 50.

A total of 1911 lineups were evaluated in the ten experiments. Each person evaluated at

least 10 lineups except for experiment 9 where three lineups were evaluated by each person. A test plot was shown to each observer and the feedback received from that plot is used to examine the data quality or process payment. But the data received for the test lineup is not included in the analysis. In some cases the demographic information were not provided by the participants. Also, for some ip address, the actual geographical locations could not be retrieved. This resulted some missing demographic information.

3.4.2 Demographic Factors

Proportions of correct responses and natural logarithms of average time taken to evaluate each lineup are computed for different demographic factor levels. Their distributions are shown using boxplots in Figure 3.4. Averages of these distributions are represented by dots inside the boxplots. The youngest age group (18-25) took least average time to evaluate a lineup. People from India took more time than others. People who have only a high school degree took less time and this is related to what we have seen for young age group. Averages of log time taken by male and female participants look similar but there are some differences as we see in Table 3.2.

The distribution of time taken to evaluate a lineup is positively skewed. But in log scale it appears to be symmetric for all the demographic factor variables as we see in Figure 3.4. Mean and medians are very similar as well. This allows us to fit linear mixed model to the log time taken with normal error structure.

We observed some differences in median proportion of correct responses for different demographic factor levels. But the variability of proportion correct responses are huge as per design of the experiments since there were some very easy as well as extremely difficult lineups. Some of the lineups were not expected to be evaluated correctly as the null plots in those lineups were showing more signal than the actual data plot itself. In those scenarios, one may not reject the null hypothesis. We call these difficult lineups. Some of the lineups were so easy that data plot is detectable at a first glance. These were expected to be evaluated correctly 100% of the time. This is why we see the boxes range from 0 to 1 in most of the factor levels. Unlike median, fewer differences are observed in the average proportion of correct responses for different factor levels.

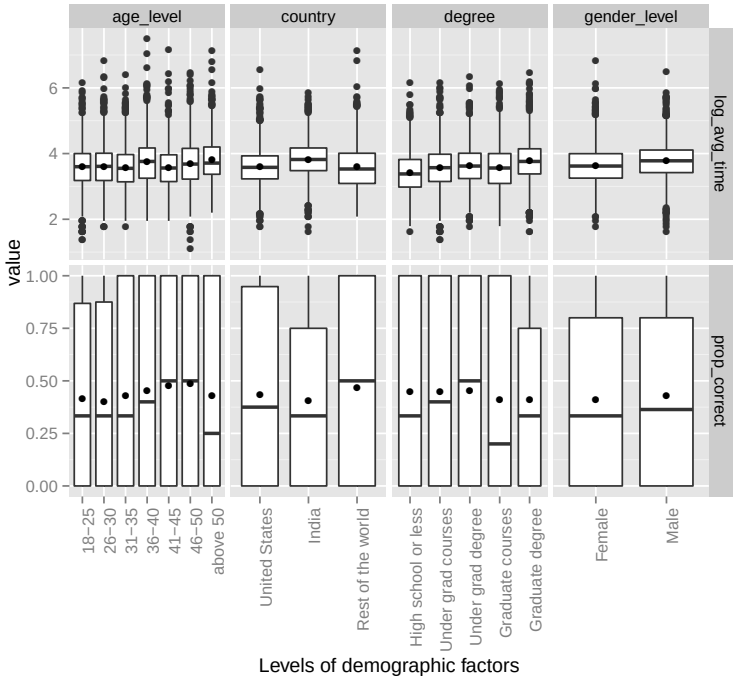


Figure 3.4 Boxplots of average log time taken and proportion correct responses of all the lineups plotted for each demographic factor levels. The dots inside the boxes represent means. Some differences in means of various demographic factors are observed. Variability in proportion correct indicates large variability in lineup difficulties.

Model (3.1) is fitted to the data with fixed effect factors such as age, country, education and gender. To estimate the overall factor main effect, a reduced model is fitted removing that factor from Model (3.1). Analysis of variance (ANOVA) results shown in Table 3.3 suggest that all the factor variables are significantly different in describing the the probability of correct response except gender. Gender does not show any significant difference in performance. Similarly, ANOVA results are obtained by fitting Model (3.2) to the data of log time taken and are shown in Table 3.3 as well. All the factor variables are significantly different for log time taken including gender.

Table 3.3 Anova of full model with all the demographic factors vs reduced model with removing respective factor variable. Gender does not have any effect on probability of correct response.

Model		AIC	BIC	logLik	Chisq	Chi.Df	p-value
Proportion Correct	Full	23822.00	23943.00	-11896.00			
	Reduced						
	Age	23835.22	23907.93	-11908.61	25.50	6	<0.001
	Country	24099.57	24204.72	-12036.78	281.85	2	<0.001
	Education	23837.48	23926.34	-11907.74	23.77	4	<0.001
	Gender	23821.88	23934.98	-11896.94	2.17	1	0.140
Log Time	Full	51904.00	52034.00	-25936.00			
	Reduced						
	Age	52435.63	52516.41	-26207.81	543.18	6	<0.001
	Country	52849.92	52963.15	-26410.96	949.47	2	<0.001
	Education	52012.68	52109.61	-25994.34	116.23	4	<0.001
	Gender	51970.57	52091.74	-25970.28	68.12	1	<0.001

To further illustrate how the individual factor levels differ from each other the detailed results of Models (3.1) and (3.2) are shown in Table 3.4. For parameter estimation the first factor levels shown in the Table are used as the base line. Demographic factor levels have significant effects on log time taken to evaluate a lineup. But not all the levels are significant for the probability of correct responses.

Age group 36-40 is significantly different than the base line age group of 18-25 year olds. The other age levels are similar in explaining the probability of correct response. For country, the rest of the world is different from USA but India appears to be similar to USA. We see from Table 3.2 that most of the participants (about 90%) are from India and USA. The rest of the 10% data are from 76 different countries. This diversity in the rest of the world may make

Table 3.4 Parameter estimates of Models (3.2) and (3.1) fitted for average log time taken and probability of correct lineup evaluations respectively. For time taken all the demographic factors are significant. For probability of correct response age group 36-40, rest of the world and graduate degree are significantly different. For gender no difference in performance is observed. Lineup variability is estimated to be very large for Model (3.1).

Demographic		Model (3.2) Log Time				Model (3.1) Proportion Correct			
Factor	Level	Est	SE	Zval	p-value	Est	SE	Zval	p-value
Fixed Effect									
	μ	3.360	0.013	249.21	<0.001	-0.683	0.071	-9.64	<0.001
Age (α)	18-25	0.000	—	—	—	—	—	—	—
	26-30	0.058	0.013	4.50	<0.001	0.062	0.049	1.27	0.206
	31-35	0.068	0.014	4.72	<0.001	0.115	0.055	2.08	0.038
	36-40	0.231	0.016	14.05	<0.001	0.310	0.063	4.93	<0.001
	41-45	0.176	0.021	8.56	<0.001	0.158	0.081	1.96	0.050
	46-50	0.272	0.024	11.29	<0.001	0.141	0.096	1.47	0.143
	above 50	0.352	0.018	19.19	<0.001	0.147	0.071	2.06	0.039
Country(κ)	United States	0.000	—	—	—	—	—	—	—
	India	0.101	0.011	9.11	<0.001	0.058	0.043	1.33	0.183
	Rest of world	-0.129	0.009	-13.82	<0.001	0.185	0.035	5.22	<0.001
Education(τ)	High school or less	0.000	—	—	—	—	—	—	—
	Under grad courses	0.042	0.013	3.25	0.0011	-0.083	0.050	-1.65	0.098
	Under grad degree	-0.037	0.012	-3.21	0.0013	-0.044	0.045	-0.97	0.331
	Graduate courses	0.117	0.013	9.12	<0.001	0.070	0.050	1.42	0.157
	Graduate degree	0.046	0.011	4.12	<0.001	0.182	0.043	4.22	<0.001
Gender (γ)	Female	0.000	—	—	—	—	—	—	—
	Male	0.078	0.009	8.26	<0.001	0.055	0.036	1.50	0.133
Random Effect									
	lineup(σ_ℓ)	0.082	0.287			5.259	2.293		
	Error(σ)	0.479	0.692						

it different from India and USA.

The graduate degree holders are significantly different as well. The positive parameter estimate indicates that they perform better and the probability of correct response is higher than other education levels. There is no significant difference between male and female performances.

Notice that lineup specific variance estimate is 5.259 with a standard error of 2.293 which is much higher than the other significant parameter estimates in Model (3.1). This indicates that the major and most important factor affecting the probability of correct response is lineup difficulty. This is also evident from the large variability in the proportion correct in Figure 3.4. Because of this huge impact of lineup specific variance, the practical impact of other factor variables on the probability of correct response is in fact very small. We illustrate this with the following example of graduate degree.

While some of the demographic factors are strongly statistically significant, the main source of variation in proportion correct is the lineup difficulty. For example, let's examine the effect of graduate degree. To see just how large the effect is, we examine the change in proportion correct for a (hypothetical) 18-25 year old female in the United States, with a graduate degree as compared to a high school degree, for an average difficulty lineup (random effect = 0). Plugging in the relevant quantities to the fitted model gives a difference equal to:

$$\frac{\exp(-0.683+0.182)}{1+\exp(-0.683+0.182)} - \frac{\exp(-0.683)}{1+\exp(-0.683)} = 0.377 - 0.336 = 0.041.$$

The person with a high school education on average picks the data plot in 33.5% of lineups having average difficulty, as compare to 37.6% if they have a graduate degree. This difference is reduced to 2% for a lineup with one standard deviation order of magnitude difference in difficulty. For two standard deviations it further reduces to 0.3%. Thus although there is statistically significant difference in proportion correct for some demographic factors, these are not practically significant differences. Figure 3.5 illustrates this example showing fitted models for a US 18-25 female with either a high school education or a graduate degree. Similar calculations show the same negligible impact of age level 36-40 (0.0533 at most) and country (0.0424 at most) on the probability of correct response. Thus even though some of the demographic factors are statistically significant, practically, demographics do not substantially influence the

results.

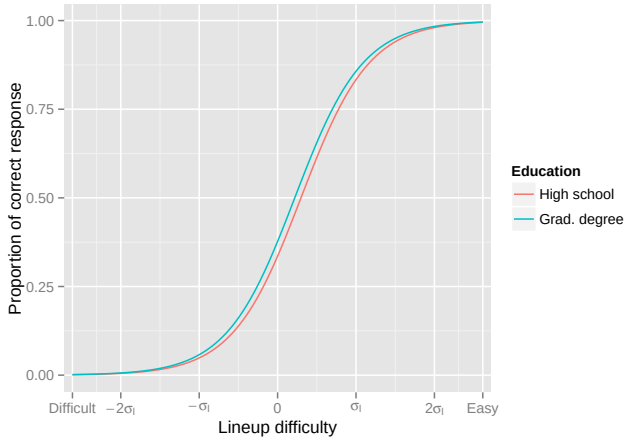


Figure 3.5 Proportion of correct responses due to graduate degree as compared to high school degree for an 18-25 year old female in the United States. Even though graduate degree is statistically significant, the largest difference in proportion correct is 0.045 which is very negligible. The difference diminishes as we move away one or two standard deviations ($\sigma_\ell = 2.293$) of lineup variability.

3.4.3 Learning Trend

Models (3.3) and (3.5) are fitted to the data from experiment 5, 6, and 7 separately. As an alternative to Model (3.3) we examined a reduced model with attempt as a continuous covariate. They did not appear to be significantly different with a p-value of 0.856. But we consider the bigger Model (3.3) with attempt as a factor variable with 10 levels to examine the effect of each level. Attempt 1 is considered to be base level while fitting the model.

Table 3.5 presents the parameter estimates and p-values of fixed effect estimates of Model (3.3). The larger p-values suggest that none of the levels of attempt (α_2 through α_{10}) are significant at %1 significance level. Moreover, some of the estimates are positive and some are negative and they show up seemingly in random order suggesting later attempts not necessarily have improved. This indicates that there may be no learning effect on the probability of correct evaluations.

The variance of random slope for attempt (σ_a^2) is estimated to be very small compared to other random effects except for subject variability (σ_u^2) for experiment 7. This suggests that most of the variability for attempt is accounted for by the fixed effect factor attempt.

Table 3.5 Parameter estimates of Model (3.3) fitted for probability of correct lineup evaluation. None of the fixed factor effects of attempt (α_2 through α_{10}) are significantly different from the first attempt α_1 at %1 level in all three experiments 5, 6 and 7. For experiment 7 subject specific variation is very small on the other hand lineup variance is much higher compared to the other two experiments.

	Experiment 5				Experiment 6				Experiment 7			
Effect	Est	SE	Zval	p-value	Est	SE	Zval	p-value	Est	SE	Zval	p-value
Fixed												
μ	-1.304	0.179	-7.28	<0.001	-0.220	0.147	-1.50	0.134	-1.737	0.481	-3.61	<0.001
α_1	0.000	—	—	—	0.000	—	—	—	0.000	—	—	—
α_2	0.270	0.219	1.24	0.217	0.262	0.158	1.66	0.098	-0.456	0.385	-1.18	0.237
α_3	-0.178	0.226	-0.79	0.432	0.342	0.157	2.18	0.029	-0.105	0.386	-0.27	0.786
α_4	0.083	0.224	0.37	0.712	0.358	0.159	2.26	0.024	-0.378	0.381	-0.99	0.322
α_5	0.298	0.224	1.33	0.183	0.376	0.159	2.36	0.018	-0.107	0.385	-0.28	0.781
α_6	0.042	0.231	0.18	0.857	0.246	0.158	1.56	0.120	0.026	0.407	0.06	0.949
α_7	0.283	0.230	1.23	0.217	0.160	0.159	1.01	0.314	0.057	0.401	0.14	0.886
α_8	-0.045	0.233	-0.19	0.847	0.341	0.160	2.13	0.033	-0.003	0.394	-0.01	0.994
α_9	-0.195	0.232	-0.84	0.400	0.378	0.160	2.36	0.018	0.204	0.436	0.47	0.639
α_{10}	0.513	0.228	2.25	0.024	0.192	0.163	1.18	0.238	-0.213	0.432	-0.49	0.622
Random												
σ_a^2	<0.001	0.017			0.001	0.034			0.027	0.163		
σ_u^2	0.720	0.848			0.815	0.903			<0.001	<0.001		
σ_ℓ^2	2.178	1.476			2.009	1.418			10.980	3.314		

For all three experiments the lineup variabilities (σ_l^2) were estimated to be very large making the practical impact of other factor levels even more negligible. Some of the lineups were very easy and some were really difficult in experiment 7. The overall average probability for picking out the data plot from a lineup (μ) is significant for both experiments 5 and 7. But for experiment 6 it is not significant suggesting that experiment 6 lineups were difficult compared to the other experimental lineups. This shows a feature of the experimental designs where a mixture of difficult and easy lineups were included within the experiments as well as between the experiments. For easy lineups there may be small chance to improve performances. But for harder lineups, the scope to improve performances is large. Since for none of these experiments improvement in performances is observed, it is important in the sense that performances did not improve over attempts in both difficult and easy lineup situations.

To visualize how the performance in correct response improves over successive attempts, we fitted Model (3.3) excluding the covariates related to attempt from the model and computed the residuals. Least square regression lines were fitted through the subject specific residuals as shown in Figure 3.6. Two important features were observed; one is subject specific variability

and the other is random slope with attempts which indicates subjects specific learning trend. Some subjects show improvement over time and some show the decrease in performance.

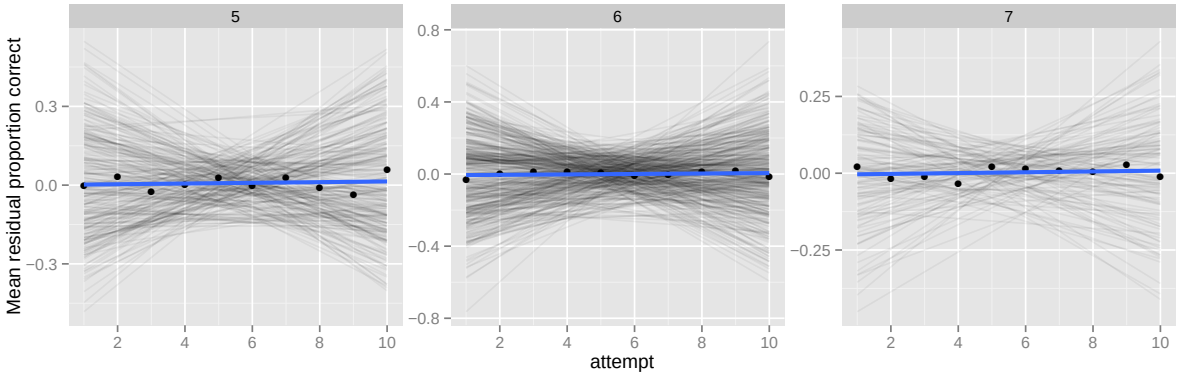


Figure 3.6 Least square lines fitted through the subject specific residual proportion correct obtained from Model (3.3) fitted without attempt are plotted against attempt. Subject specific positive and negative slopes are observed. Mean residuals are shown as dots and least square regression lines fitted through the points show no overall learning trend in each of the three experiments.

The averages of these residuals for each of the attempts are shown as dots in Figure 3.6. Least square linear regression lines are fitted through the points for each of the experiments. Positive slopes over the attempts are observed but none of them is statistically significant.

Table 3.6 presents the results of Model (3.5). The parameter α for fixed effect covariate attempt is highly significant in all the experiments. The negative estimates suggest that on an average later attempt took less time for an evaluation. Even though observers did not improve the performance over attempt, they became efficient in responding faster. The parameter α_1 for first attempt is also highly significant. The positive estimates of α_1 indicates that first attempt made by an observer required much more times than other attempts. It is because for initial attempt the observer might have gone through instructions and became familiar with the experimental environment. Each page in the web site contains information about choice reason and the observers's confidence level. Also, the first page asks for observer Identification to be typed. The later pages of the web site was similar just the lineup was changed. Thus in the later attempts an observer does not need to spend any time for reading instructions. The model reflects that fact.

The lineup variance of experiment 7 is estimated as $\hat{\sigma}^2_\ell = 10.98$ from Model (3.3). But from

Table 3.6 Parameter estimates of Model (3.5) fitted for log time taken to evaluate a lineup. Both fixed effect parameters of Attempt (α_1 and α) are highly significant for all three experiments 5, 6 and 7.

	Experiment 5				Experiment 6				Experiment 7			
Effect	Est	SE	Zval	p-value	Est	SE	Zval	p-value	Est	SE	Zval	p-value
Fixed												
μ	3.817	0.039	97.38	<0.001	3.901	0.033	118.19	<0.001	3.731	0.054	69.04	<0.001
α_1	0.326	0.035	9.35	<0.001	0.335	0.029	11.40	<0.001	0.280	0.050	5.63	<0.001
α	-0.038	0.004	-9.30	<0.001	-0.039	0.004	-10.19	<0.001	-0.029	0.007	-4.22	<0.001
Random												
σ_a^2	0.001	0.027			0.002	0.045			0.002	0.049		
σ_u^2	0.259	0.509			0.245	0.495			0.134	0.366		
σ_ℓ^2	0.008	0.091			0.040	0.199			0.055	0.235		
σ^2	0.211	0.460			0.251	0.501			0.206	0.454		

Model (3.5) it is estimated to be much smaller in all these experiments. This suggests that the harder lineups not necessarily took more time than easier lineups. The observers spent enough times to evaluate a easy lineup and for difficult lineup they might just give up at some point and provided the feedback.

Model (3.5) was compared to couple of other alternatives. A bigger model with attempt as a factor was considered but it was not significantly different with p value 0.236. A reduced model without the first attempt was also considered but it was significantly different with p-value < 0.001.

To visualize how the time taken reduces over the successive attempts, we fitted Model (3.5) excluding the covariate attempt from the model and computed the residuals. Least square regression lines are fitted through the subject specific residuals. Subject specific slopes are much different in each of the three experiments as we see in Figure 3.7. Some subjects improved over attempts by taking less time in the later attempts while others got worse. The averages of these residuals for each attempt are plotted as dots. Least square regression lines are fitted to these points excluding the first attempt since for first attempt we fitted an indicator covariate. The downward trends are evident in the plots. All the slopes are highly significant. As expected we observed large positive residuals for each of the experiments for first attempt.

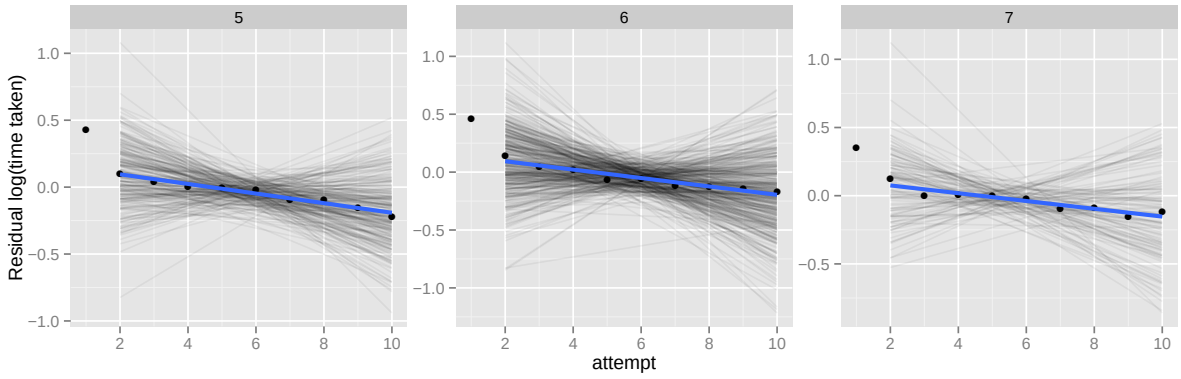


Figure 3.7 Least square regression lines fitted through the subject specific residuals obtained by fitting Model (3.5) without covariate attempt. Differences in subject specific slopes are observed. Some of the subjects did worse over successive attempts while others did better. Averages of these residuals are plotted as dots and least square regression lines are fitted to obtain overall trends. For all the three experiments the overall downward slopes are statistically significant which indicates that MTurk workers take less time as they progress through their attempts.

3.4.4 Location Effect

A total of 111 subjects were recruited to evaluate lineups designed to investigate the location effect of the actual data plot in the lineup as described in Section 3.3.1.2. Each subject evaluated two lineups; one for Interaction effect and the other for Genotype. In total there were 222 feedbacks or responses on 50 lineups. The data on the test lineup were excluded from the analysis.

The proportion of correct responses for each data plot location is shown in Figure 3.8 colored by null sets. We observe some variability of performance for different null sets even though same data plot was used for all these null sets. This may happen when for some set of null plots, one null plot appears to be very similar to the actual plot. In another set of null plots this may not happen making some lineups easier than others even though the actual data plot is the same. This pattern is evident in the figure as we see proportion correct for null plot 5 is consistently above the null plot 1 for each of the locations. A test for differences in average performances of null sets shows significant difference only for null set 3 of interaction lineup. The rest of the null sets don't show any differences at 1% significance level.

The overall average proportion of correct responses for each location are shown using dashed

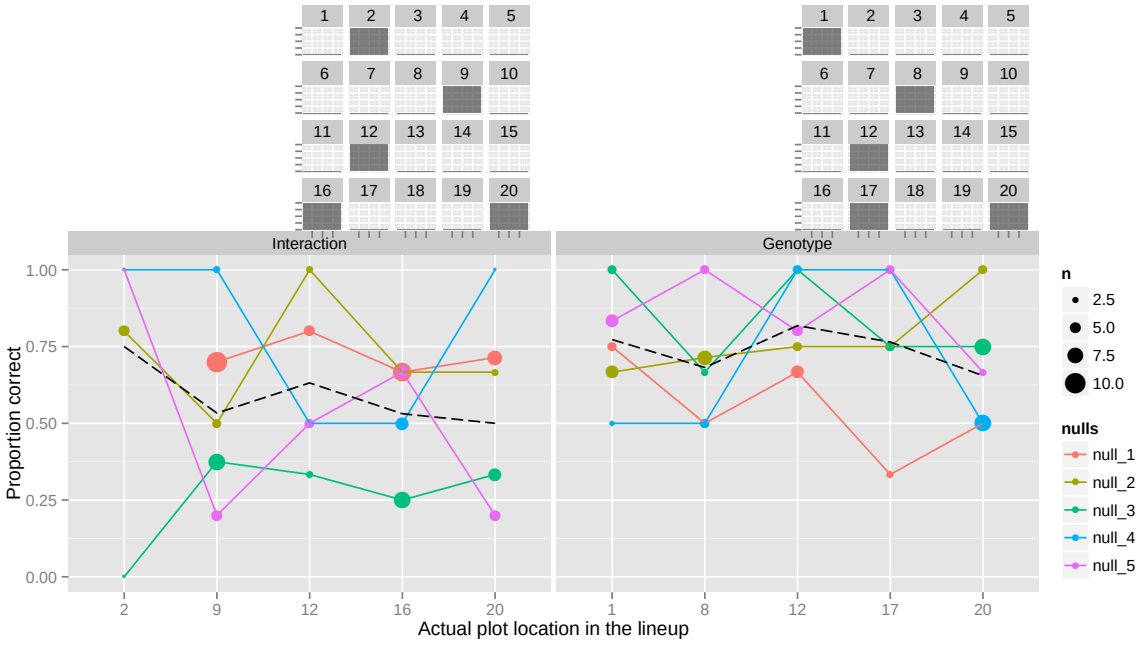


Figure 3.8 Location of data plot in the lineup and proportion correct for both Interaction and Genotype effect. Each colored line represents a null set and the size of the dots represents number of responses. The overall average proportions are shown by dashed line. The actual data plot locations are shaded grey on the top panels to demonstrate their relative positions on a lineup.

lines in Figure 3.8. There is variability in performances for each null set depending on the locations. But the overall proportion is not varying much for different locations.

The sizes of the dots in Figure 3.8 represents the number of responses obtained for that location and null set. For some locations we have as many as 10 responses. For location 1, we did not have any response for null set 1 in one of the interaction lineups. We observe larger variability for interaction lineups as compared to Genotype lineups. It is because some of the interaction lineups were evaluated only once as we see for null plot 3 and 5 for Interaction effect. This leads to the extreme proportion of correct since one response will either be correct or wrong. The larger variability observed for Interaction lineups can be controlled at some extent by evaluating them multiple times.

We fitted Model (3.7) to test if the mean performance vectors are similar for different locations. For this we use *anova()* function of *stats* package of R Core Team (2012). The results are shown in Table 3.7. The p-values for both Interaction and Genotype effect are much bigger than the conventional threshold of 0.05. This suggest that there may be no difference in location.

Table 3.7 The results obtained by fitting MANOVA Model (3.7).

Location				Degrees of Freedom			F test
Effect	DF	Pillai	Approx. F	Numerator	Denumerator	Residual	p-value
Interaction	3	1.4783	0.7772	15	12	6	0.6821
Genotype	4	1.7796	1.1221	20	28	8	0.3824

We also examined whether the proportion of correct responses differs if the actual plot is on the outer boundary or in the inner locations. The locations 7,8,9,12,13,14 in the lineup are considered inner locations and the rest of the boundary locations are considered to be outer location. As we see in Figure 3.8, location 9, 12 are inside for Interaction effect and location 8, 12 are inside for Genotype. It does not show any differences whether the actual plot is inside or outer border of the lineup. We also fitted Model (3.7) with two locations, inner and outer, as covariate and observed no significant differences.

3.5 Conclusion

Human demographics have significant influence on performance in time taken to evaluate a lineup. Some variations among the factors are observed in terms of probability of correct evaluation. Age group 36-40, Countries other than India and United states, People who have a graduate degree are significantly different. But their practical impact on probability to correctly evaluate a lineup is very negligible and in some cases it diminishes as the lineup difficulties increase or decrease. Gender does not have any significant effect on performance. Thus there may be differences in time taken to evaluate a lineup for different human demographics but the practical impact of demographics on the performance is very negligible. This result is very important for the power of visual test as it demonstrates the robustness of the test for different human factors.

Individual learning trend is observed in both time taken and observer performance. Some individuals improved the performances while others showed decrease in their performances over attempts. But the overall performance of the observers does not increase through successive attempts while evaluating multiple lineups. This result suggests that the power estimated for visual inference using human subject experiment is robust and may not change if those participants are allowed to give feedback again. The skill in evaluating the lineup in shorter time gets improved over successive evaluations. The earlier evaluations take significantly longer time than the later evaluations. It suggests that the skilled person may only do it faster.

The simulation experiment reveals that there is no significant effect of location of actual data plot in the lineup. This is important as the visual statistical inference procedure suggests that the data plot be placed at random anywhere in the lineup. This paper suggests that any random place in a lineup is as good as other places in the lineup. Even though there are variations on the performance depending on different null sets, their impact on probability to correctly evaluate a lineup is very negligible.

The subjects participating this study may not necessarily know about statistical graphics. The numerous pilot studies done with more trained participants suggest that power of visual statical inference may be higher for observers who have advanced knowledge about statistical

graphics. The fact that a person with graduate degree performs better may be associated with this. But a graduate degree does not necessarily mean to have training on statistical graphics. More experiments may be needed to learn about the differences in power with trained and a non-trained observer in terms of knowledge about statistical graphics.

This investigation allowed each observer to evaluate 10 lineups assuming that it would not cause fatigue or disinterest toward the task of evaluating lineups. This assumption is made based on the pilot studies. The future research may involve doing more experiments to check if fewer or more than 10 lineups have any significant impact on performance of the observer. The learning trend in terms of time taken shows downward trends. At some point it should level of. It will be useful to know for how many lineups time taken levels off to decide how many lineups is appropriate to show.

A lineup with fewer or larger than 20 plots may yield different results. If the size of the lineup is much higher than 20, it is intuitive that there may be location effect of the data plot in the lineup. It is because, the observer may get tired of scanning and may make decision based on the partial scanning of the lineup. On the other hand fewer than 20 plots will allow observer to compare the actual plot with fewer null plots giving more chance of picking actual plot as an error not as an actual plot. These are the some of the issues that require further investigation.

Acknowledgments This work was funded in part by National Science Foundation grant DMS 1007697. All studies were conducted with approval from the Institutional Review Board IRB 10-347.

CHAPTER 4. DESIGNING TURK EXPERIMENTS FOR VISUAL STATISTICAL INFERENCE

A paper to be submitted to *Journal of Statistical Software*

Mahbubul Majumder, Heike Hofmann, Dianne Cook

Abstract

Human observers are needed to evaluate lineups that are used to test the significance of findings using statistical graphics. One good option is to recruit people from online workplace like Amazon Mechanical Turk (MTurk). It provides features to create online task that allow people to evaluate lineups and get paid. MTurk is designed for simple and easy tasks. The technical design of the underlying experiment for lineup evaluation may be complex and the tools available to design this from MTurk is just too simple. In this paper we present the design of MTurk experiments for lineup evaluation, provide an alternative way to conduct the survey on lineups by developing a separate web application and getting turk worker do their job from that web site. The web site is now hosted on the Iowa State University public domain and has been in use for multiple experiments (Majumder, 2013). It provides multiple features that make it convenient to get lineups evaluated by online observers and obtain data in a secure way.

Keywords: statistical graphics, lineup, Amazon Mechanical Turk, visual inference, visualization, Crowdsourcing, Human Intelligence.

4.1 Introduction

There have been some advancements in visual statistical inference since the concept was first introduced by Buja et al. (2009). They proposed lineup protocol that allows testing the significance of findings using statistical graphics. In visual statistical inference the test statistics is a plot of the observed data. This plot is placed randomly in a layout of plots called lineup. The rest of the plots, called null plots, in the lineups are generated using the data simulated from the model specified by the null hypothesis. A human observer is asked to evaluate the lineup. If the observer can detect the actual plot in the lineup, the null hypothesis is rejected.

A lineup is shown in Figure 4.1 where the observed scatterplot is displayed with a least square line overlaid. Can you find which of these plots show the steepest slope? The null plots are generated based on the hypothesis that the slope is zero. If the observed scatterplot is detected, it indicates that observed plot is different from the null plots and hence provides evidence against the null hypothesis.

Majumder et al. (2013b) further developed visual statistical inference by refining the terminologies, presenting the ways to compute p-values associated with the visual test and providing the methods of obtaining the power of the test. It is revealed that the power of visual test can be as good as that of the best available conventional test and in some scenarios even better. This work establishes the validity of lineup protocol to be used as a tool for statistical testing. In the situations where no conventional test exists, visual inference can be the only inferential procedure that does not compromise a lot on power because of its non-parametric nature and few assumptions.

These developments open up a new area of statistical research where lineups need to be evaluated by human observer. For small scale or day to day research it is possible to have people around to get the lineups evaluated. But sometimes it is needed to have many observers to evaluate lineups. For example, to assess the power of the visual test for different visual test statistics Hofmann et al. (2012) recruited human observers to evaluate lineups. Roy Chowdhury et al. (2012) used lineup protocol in practical applications. Other reasons to evaluate lineups may be to present the results of the conventional test with visual tools such as lineup.

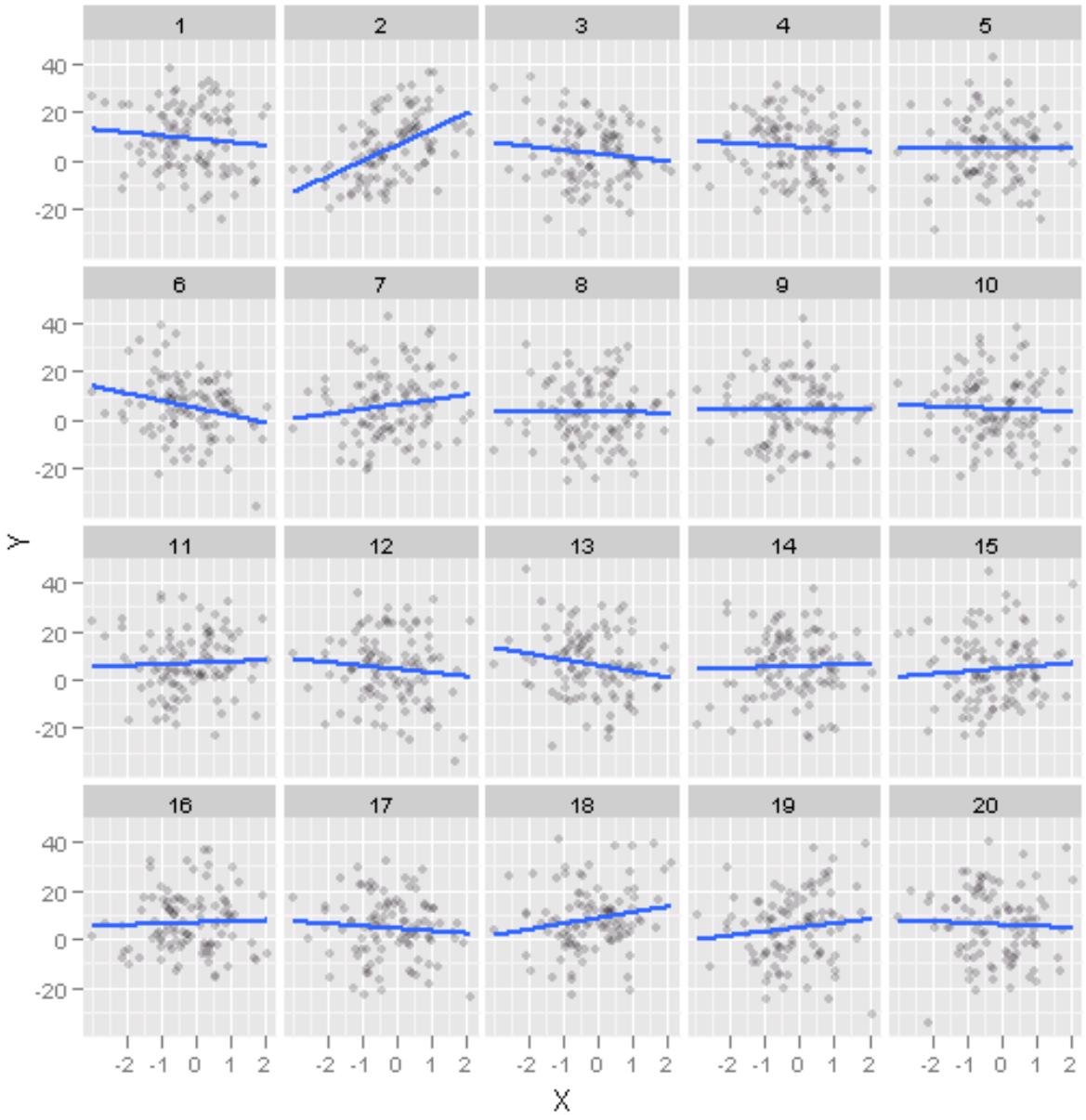


Figure 4.1 A lineup of 20 scatter plots with least square line overlaid. Which of these plots shows the steepest slope? Answer to this question can be found at the end of conclusion.

The power of visual test can be obtained theoretically under some assumptions on individual behavior which was evident from experimental studies (Majumder et al., 2013b). In general it is hard to obtain explicitly with a mathematical formula since it is very much dependent on human observation. One approach to estimate the power is to recruit observers to evaluate lineups generated with some known effect sizes. The proportion of correct evaluation can be used as an estimate of the power for that effect size. Individual differences in the abilities of correct evaluations of lineup were observed. Thus it is desirable to get observers as diverse as possible.

This poses a new challenge to researcher to recruit people from a diverse pool of population. Cost, time, data qualities and convenience are some of the issues that need to be dealt with. Fortunately, we can use the services of Amazon Mechanical Turk web site for this which is discussed in the following section.

4.1.1 Amazon Mechanical Turk (MTurk)

Amazon (2010) Mechanical Turk or MTurk is an online work place where people from around the world can perform tasks and get paid. Usually tasks are very simple and no specialized training is necessary. Being a human is the main requirement. Tasks are designed for anyone to do but some tasks may require that workers satisfy some skill level depending on the recruiters' need. The tasks are designed to be done in a very short time. Humans are still better than computers in performing these types of tasks. The the amount of money paid for each task is very small as well. Figure 4.2 shows an MTurk task where an observer is asked to select an option based on a picture.

It is very cheap and reliable to recruit people from MTurk and the results can be obtained very fast. It allows distributing works to the thousands of workers around the world. The another benefit is that a very diverse pool of subjects can be recruited which is otherwise very hard to obtain for a study. The researchers can easily filter the workers based on their experimental design, such as recruiting people only from a specific geographical location or a group of people who satisfy certain criteria etc. The recruiter can decide who they pay or not. Workers have to satisfy the task requirement to ensure payment. But at the end it is the

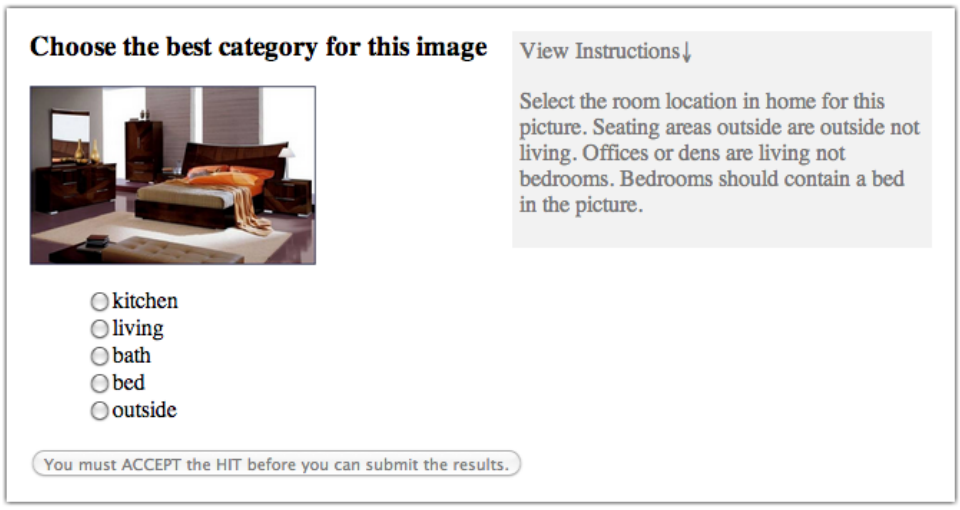


Figure 4.2 An example of amazon mechanical turk task. Tasks are usually very simple and designed for human evaluations. With each task, simple instructions are given for workers to follow. The workers first accept the task before submitting their response.

recruiter who has the final say. Usually recruiters pay promptly after the task has been done properly and thats why MTurk is very popular among the online job seeker.

Because of its convenience it is getting popular for scientific research study as well. In comparison with a lab study Suri and Watts (2010) performed the same study using MTurk and demonstrated that their study results are as good as the lab study results even though MTurk study required less time and cost while provided more convenience. Majumder et al. (2013b) recruited people from MTurk for their simulation study in estimating the power of visual statistical inference. They have done numerous pilot studies in lab before doing actual MTurk study and found similar results. Mason and Suri (2012) explains various features of MTurk and describes how it can be used as part of human behavioral study.

The simplicity of the MTurk task is the main factor for its popularity. Figure 4.2 shows how simple an MTurk task could be. It is possible to get a lineup evaluated by creating such a simple task. We just need to replace the picture in Figure 4.2 with the lineup in Figure 4.1 and change the answering options. But some times we may need more than one lineups to be evaluated by an observer. We may need to show a random sample of lineups from a pool of many lineups automatically. The questions the observer would answer while examining the

lineup can be different based on different lineups. Moreover, It is convenient if the lineup is clickable so that selection of plot can be made by the mouse click. To create an MTurk task that has all these flexibilities, a web application needs to be created. The following section discusses how the system would work.

4.1.2 Getting Turk Workforce

Once a web application is created, there are two options to run it; one is to run it inside MTurk system using their API and the other is to develop a new web site separate from MTurk. For additional control we picked the second option and planned to separate this application from MTurk system and designed our own web site to run the application. It enables us to display the lineups to the observers with a lot of flexibilities. As an added benefit the data can be directly saved in a local database server instead of getting it from MTurk.

Figure 4.3 shows how the plan works. First an MTurk task is created for the workers to review and decide whether they want to do the task based on the parameters like payment amount, estimated time the task may take to finish and how hard the task seems to be etc. Workers are informed that the task has to be done outside the MTurk system from another web site. Once the workers accept the task they are redirected to our web site where multiple lineups are shown for evaluations. After the required number of evaluations has been received, a code is generated which the workers need to submit back in MTurk system to complete the task. The code is matched with the code in our database to process payment through MTurk system.

MTurk provides workers who usually do the simple tasks available for them. Thus, for lineups to be evaluated a simple task needs to be created through MTurk. This may be a challenge for the researchers who want to make statistical decisions based on lineups or need to study the power of visual tests. They need a system that can provide an easy solution to this challenge while giving them flexibilities how they want to present the lineups. This paper provides a complete solution to this. It is organized the way a researcher may work with the experiments related to lineups and like to get feedback data from the observer. Section 4.2 presents the detailed description of an experiment with lineups and things to consider while creating a turk

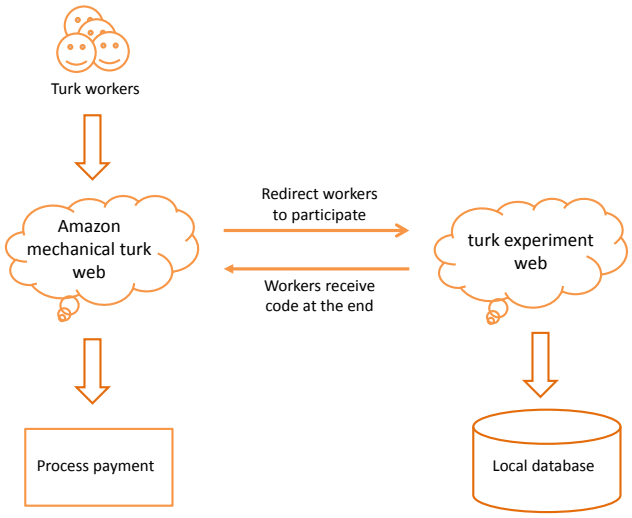


Figure 4.3 Amazon Mechanical Turk workflow shows how data are collected through turk experiment web and payment is processed through MTurk system.

task. Section 4.3 presents the design of an web application to get lineups evaluated with additional flexibilities and options that may not be possible through MTurk system. Section 4.4 describes how to create an MTurk task, mange workers and process payments. Finally Section 4.5 presents some data obtained from various experiments done through the web application and discusses about the quality of the data.

4.2 Experiment Design

If a lineup is created from the actual observed data, the web application presented in Section 4.3 can be used to get it evaluated. In that case we do not need to design a simulation experiment. This section will be useful when a simulation experiment is needed to examine different visual test statistics and compare the power (Majumder et al., 2013b). To examine the effectiveness of certain plots in displaying the data one may want to design an experiment with lineup (Hofmann et al., 2012). In any cases, the two major considerations are the choice of parameters to simulate lineups and number of people needed to evaluate a lineup.

4.2.1 Selecting Parameter to Simulate Lineup

In any experimental design we need to have some control. For a simulation experiment with lineup, this refers to the parameters related to a lineup. This depends on what the purpose of the experiment is. For example, to compare the power of a parametric test one may fix the parameter specified by alternative hypothesis and generate data from that model. This data can be considered as the observed data. A lineup can be created by placing this observed data plot in a layout where rest of the data plots come from the model specified by the null model (Majumder et al., 2013b).

In addition to the hypothesized parameter of interest, other parameters that may need to be fixed include sample sizes, variability in the data, error structure etc. These parameters produce the effect sizes and are responsible for any detectable signal in the lineup. Thus for a simulation study a range of effects sizes need to be considered. It is a little bit tricky to decide what exact effect sizes are appropriate. For example in Majumder et al. (2013b) the parameters were chosen so that they produce a smoother power curve of a conventional test. But in general this depends on what the researchers may want to test and control in an experimental study.

4.2.2 Procedure to Simulate Data Plot

In the simulation study the very first step is to generate a random sample of data from the model of interest with pre-specified parameters. We call this data set observed data. Every time we do that we get a new set of observed data which may produce an estimated parameter very different than the actual parameter specified to generate the data. But it is important to have an observed data closely representing the parameters chosen since other null plots will be compared to this data plot in a lineup. One solution to this is to take some replications of lineups for the same effect size.

While the effect of the natural variability of the observed data set can be controlled by taking the replication of few samples, it is desirable to study whether we can reduce this variability further to some extent. For this we pick the example of a simple linear regression model and examines how the data plot can be generated that best represents the specified parameter.

Consider a linear regression model

$$Y_i = \beta_0 + \beta X + \epsilon_i \tag{4.1}$$

where $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, 2, .., n$. The covariate X can be continuous or discrete. As discussed in Section 4.2.1 we specify the parameters sample size n=300, regression slope β=3 and error standard deviation σ=12 to simulate observed data.

To make sure that the observed data set truly represents the Model (4.1) we suggest three approaches. One is Kolmogorov test statistic approach and the other two approaches are quantiles of p-values and closeness of estimated parameters to the true parameters. To illustrate the three approaches we used Model (4.1) as an example but the idea can be extended to any situation. The three approaches are discussed below.

Kolmogorov test statistic approach: In this approach we simulate 1000 data sets from Model (4.1) and obtain Kolmogorov test statistic for each set of data as below.

$$D_n = \sup_x |F_n(x) - F(x)|$$

where $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{X_i \leq x}$ be the empirical distribution function of fitted residuals, $I_{X_i \leq x}$ be the indicator function equal to 1 if $X_i \leq x$ and equal to 0 otherwise and $F(x)$ be the cumulative function of normal with mean zero(0) and variance σ^2. We keep the data set that has minimum value for Kolmogorov test statistic since for this data set $F_n(x)$ should be the closest to the desired normal model.

Quantiles of p-value approach: For a simulated observed data set we can obtain the p-value associated with testing $H_0 : \beta = 0$. The distribution of p-value is uniform under null hypothesis but under alternative it has a right skewed distribution. Figure 4.4 shows the distribution of p-values for sample size n=300, regression slope β=3 and error standard deviation σ=12. If we generate a data set randomly there is a 21% chance that the data will show a p-value of 0.25 or more even though we generated data set with non-zero $\beta = 3$. We need to make sure that the simulated data does not come from this extreme end. Additionally,

we want some replications so that this extreme effect can be controlled at some level. For this example we intend to pick three replications.

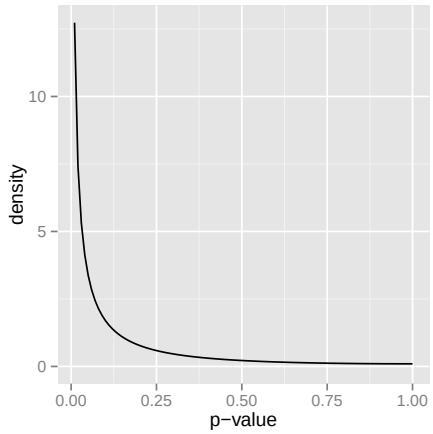


Figure 4.4 Distribution of p-values under alternative hypothesis ($H_1 : \beta$=3) for sample size n=300 and error standard deviation σ=12.

In this approach we generate 1000 data sets from the model of interest and obtain p-values for each data set after fitting the same model. We construct 3 blocks of p-values such as (0.0-q_{33}), (q_{33}-q_{66}), (q_{66}-1) where q_i is the ith percentile in the distribution of the p-values. We randomly select three data sets that have corresponding p-values in the above quantile range. Notice that if we like to have three replications of plots, we will have observed data sets with smaller p-values as the distribution is highly right skewed as well as the 33th and 66th percentiles of p-values are 0.0101 and 0.077 respectively. So, another option may be to take the mid points of those quantile ranges and pick the data sets that produce those p-values.

Closeness of estimated parameters: When we simulate a data set from the Model (4.1) and fit the model to the data set again, we do not get the parameters estimates same as what we fixed while simulation. For example, with a simulated observed data we obtained a parameter estimates of 0.5 for true $\beta = 3$. We want a data set where the parameter is closed to our fixed slope $\beta = 3$. In this approach we generate 1000 data sets and pick the data set that shows most close estimates of parameters compared to true parameter values. We do this three times to obtain three replications of the data.

All of these three approaches discussed above have some problems. Kolmogorov method does not necessarily make sure that estimated parameters are close to the true parameters. On the other hand closeness of estimated parameters does not make sure that p-value is small enough when we should reject the null hypothesis. Only the quantiles approach seems reasonable as it has much control over p-values and it gives similar data sets that has closest parameter estimates. So we suggest the quantiles of p-value approach to generate observed data with specific effect size.

4.2.3 Sample size estimation

For any experiment it is important to determine the sample size. It is not only related to time and cost of the experiment but also associated with the validity of the results. In simulation experiment with lineup the sample size means the number of evaluations per lineup. It is different than the number of people who evaluates lineup. If multiple lineups are evaluated by a person, sample size will be larger than number of people recruited from MTurk. For example, if a lineup needs to be evaluated 20 times, we need at least 20 persons since each person does not see the same lineup more than once. This will also enable us to get 10 lineups evaluated if all of these 20 people evaluate 10 lineups each. But all the lineups don't require same number of evaluations. For easy or difficult lineups fewer evaluations are needed.

One approach to assess number of evaluations that are required for a lineup is to have a prior idea of proportion correct responses for the that lineup. For a given proportion p we want to have margin of error (ME) to be at most 0.05. Thus we have

$$ME = 1.96\sqrt{\frac{1}{n}p(1-p)} \leq 0.05$$

which gives us the estimation of minimum sample size

$$n \geq \frac{p(1-p)}{(0.05/1.96)^2}.$$

If we do not have any prior idea of proportion of correct or power, we may rely on the power of a similar conventional test if available. This can be assumed to be the power or proportion correct for that lineup. Other ways of having an estimate of sample size can be very specific to the problem of interest and should be computed as required by the specific problem.

4.2.4 Test and Training Lineup

It is desirable that some lineups be displayed to the MTurk workers so that they can become familiar with the experimental environment before they participate the actual experiment. It is important to setup this properly. These training lineups should be easy and feedback should be provided right after the response is received. The purpose is to give them an idea how the actual task would look like. The workers may or may not opt for trying the training lineups.

Another option is to make this training mandatory to participate the experiment. The MTurk workers should have certain proportion of correct responses on these training lineups to be allowed to participate. This is often practiced in MTurk task and the workers are very familiar with this approach. In MTurk language it is called the qualification of the worker. Thus to design a MTurk task it is important to fix a standard qualification of the worker.

Other things that need to be considered include number of training lineups and how the lineups should be shown. Training should not take much longer. Two or three lineups can provide enough training. The training lineups can be randomly selected from a very small number of lineups. That may produce repetition which may help the MTurk workers to retest their skills on the lineup they might have a wrong selection before.

There should be a test lineup when the workers evaluate multiple lineups in actual experiment. This test lineup can be used to process payment and detect any unusual responses. The test lineup should be super easy to evaluate so that anyone can detect the data plot without much effort.

It is helpful if some example lineups are presented with the experimental description. Example lineup does not have to be of the size similar to the actual lineup. it can be of size three or four so that more than one lineup can be placed on a page for demonstration of the task. This helps the worker decide whether they are willing to participate the experiment at all.

4.2.5 Plan for a Turk Task

To make a good use of MTurk crowd source, it is important to split any big project into small pieces or Tasks. To present each task for the workers to do, a Human Intelligence Task or

HIT needs to be designed which includes crisp and clear instructions of the task, descriptions of specific input and output desired and payment information. To design a HIT for lineups to be evaluated we need to consider the following issues;

- **IRB approval:** Since this is a human subject experiment proper approval has to be taken from Institutional Review Board or IRB. This approval includes how the task will be performed, how the anonymity of the subjects will be maintained etc. For this we plan to have a consent form which the MTurk workers have to agree before participating the experiment. The consent form needs to be approved by IRB.

- **Lineup question:** Each lineup should be presented with a question that the worker is asked to answer. The question needs to be clear and technical words should be removed as the MTurk workers may not be aware of those terms. The question is important for lineups as well. Very common question to evaluate a lineup is "Which plot is the most different?". But for complicated studies the question may be quite different (Majumder et al., 2013a).

- **Data input:** For lineup experiments input data mainly refer to the information about lineups that are going to be presented for evaluations. Lineups may be presented in some sequence or completely at random. We need to decide whether a test lineup will be included or not in the pool of lineups going to be presented. Some other inputs may accompany with each lineup. Those include lineup question, how many lineups to be evaluated by the worker etc. Anything that the worker will see changing with each lineup is considered as data input to the worker. For different experimental needs this input may be different. For example, one may decide to let the worker know if each evaluation made is correct or wrong after the feedback has been submitted. This is an input for the worker.

- **Data output:** This means the data to be collected from the MTurk workers. That includes mainly the response on lineup question. Single or multiple responses can be requested for each lineup. Other optional data that may be collected are as follows;

1. Demographic information of the observer
2. Supplementary information such as reasons or confidence levels of the response provided
3. Location information
4. Time of the evaluation

- **Payment:** Payment plays an important role on data quality and how fast the data can be collected. It depends on how many lineups are being evaluated by a worker, how much time it may require to finish the task or how hard the task is in general. The MTurk convention is to pay as per an hourly rate of $6. But some times it may be more or less depending on the task type. For some easy tasks it may require longer time to evaluate reducing the pay rate by hour. Other tasks which are difficult but can be done very fast will have a increase payment rate.

4.3 Web Application for Turk Experiment

A web site is developed to get lineups evaluated by human observers (Majumder, 2013). It provides all the features needed for a simulation experiment with lineups but yet remains simple enough for the online workers to provide feedbacks. This section describes the technical details of the site.

The web application is built using server side scripting language PHP embedded in HTML. JavaScript is used to control the client side work flow such as preventing missing information, showing instructional messages etc. MySQL database is used to store data for dynamic presentation of lineup and recording observer feedbacks. The application also records the ip address of the observer's machine and the times at which each lineup is displayed and evaluation is received.

4.3.1 Form Design

Designing a data collection form is very critical in turk experiments. As we see from Figure 4.2 that a turk task needs to be as simple as possible. This is what the turk workers are

prepared for. Making a complex form may turn out bad and jeopardize the whole purpose of the experiment.

Keeping these in mind, two web forms are designed to collect information from the turk workers. The first form shown in Figure 4.5 collects feedback information about a single lineup. The information collected through this form is all about the lineup that includes plot number selected, reasons for selection and the confidence level of the selection. Each turk user is identified by the nick name which is mainly the turk ID. It looks simple and it is indeed simple for the turk worker to provide feedback using this form. But it is not simple in design as all the information on this form are coming from database including the question on top of the lineup. This dynamic page provides a lot of features in customizing how one may want to display the lineup.

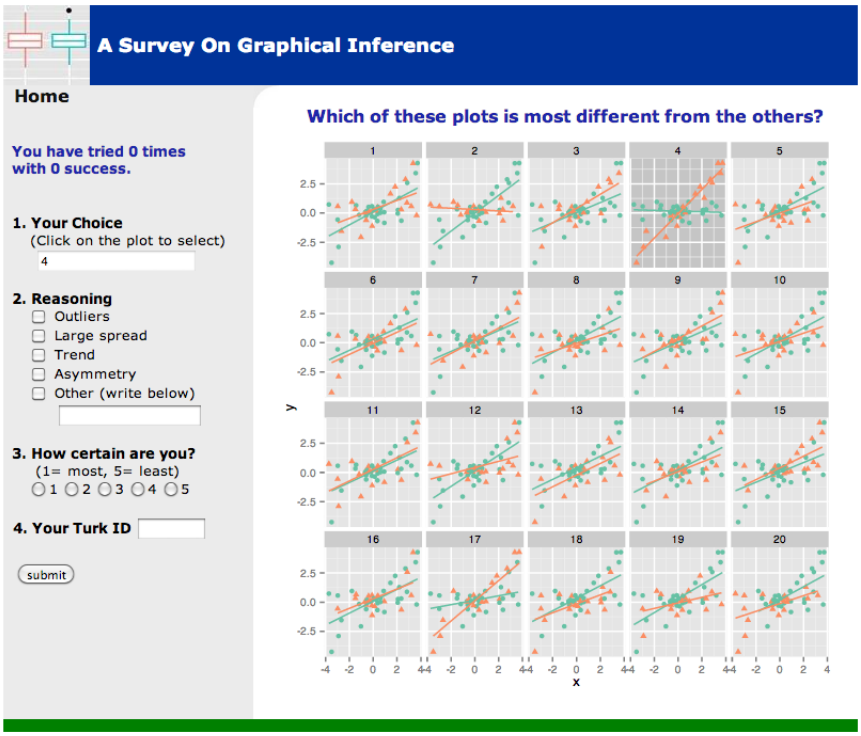


Figure 4.5 A sample data collection form. Lineups are presented at random for evaluations by the turk workers. Scalable Vector Graphics (SVG) is used so that observer can click on the lineup to pick certain plot. Once a plot is selected it gets shaded and the number appears in the choice text box.

One of the nice features of the form in Figure 4.5 is the use of Scalable Vector Graphics (SVG) for lineup which enables an observer to give feedback with the ease of just a mouse

click. If the observer changes mind he or she can click the plot again and deselect it. If needed multiple selection of plots can be allowed, order of selection and deselection can be recorded with time taken for each action. That is what we meant by saying it is complex in design.

It is possible to let workers see the statistics of their total feedbacks with number of correct evaluations. This feature can be opted out easily if not necessary. The workers have to provide their turk ID and for next evaluation they don't have to type it again. This allows the worker give feedback using only mouse click. Thats what we meant by saying it is simple for turk worker.

Once the data is submitted a feedback can be provided whether their choice was correct or wrong. This feature can also be opted out easily if not needed. The design of the feedback form is shown in Figure 4.6. This form is also used to collect demographic and educational information about the worker. After the required number of evaluations are obtained, a pass code is given as a proof of the completion of the task which is used for payment purpose later.

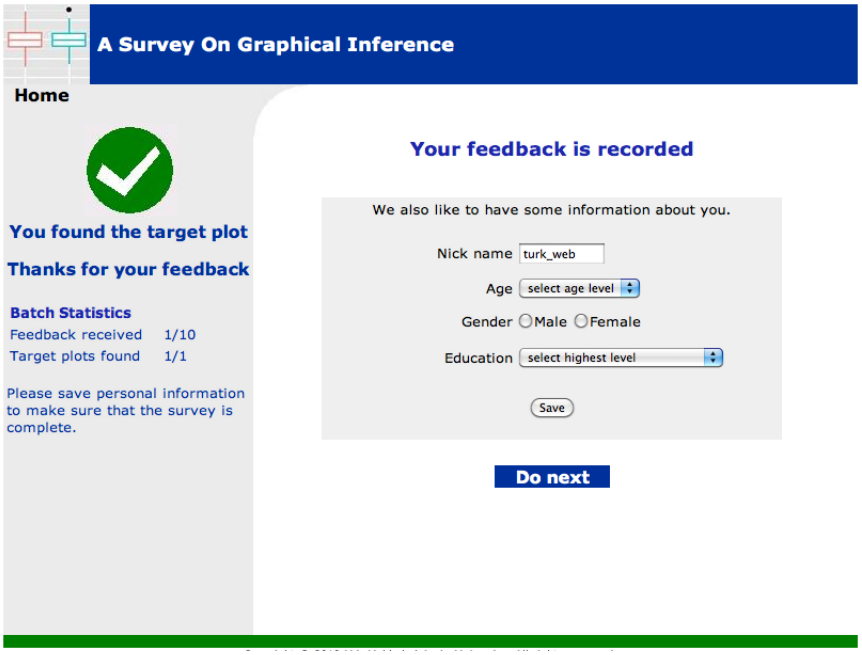


Figure 4.6 The turk workers are given feedbacks whether their evaluation for each lineup was correct or not. This works as an incentive for the worker to work more enthusiastically. To ensure the payment, the turk workers have to provide their demographic information using this form.

Each worker is shown some specific number of lineups for evaluation. These lineups are

randomly selected from a pool of lineups designed for evaluations. The algorithm of how the lineups should be selected to show and what would be the order of display is implemented in the form shown in Figure 4.5. Also there is a check for invalid or missing information which is implemented using javascript. If users try to go forward without giving any feedback they are not allowed to do that showing a warning message.

4.3.2 Database design

The design of the database to store the collected data is shown in Figure 4.7. It is designed such a way that data from many different experiments can be stored in the same database. In total five separate tables are used. Table turk_worker contains static information about each turk worker. The location information of each turk worker is stored in ip_detals table. The information in this table is collected later based on the ip address of each turk worker. Table picture_delais contains the static information about each lineup. Table feedback is a dynamic table that grows with the number of feedbacks from each turk worker. The multiple activities of the worker are recorded in turk_activity table. This table also contains the codes provided to the workers once they finish the experiment.

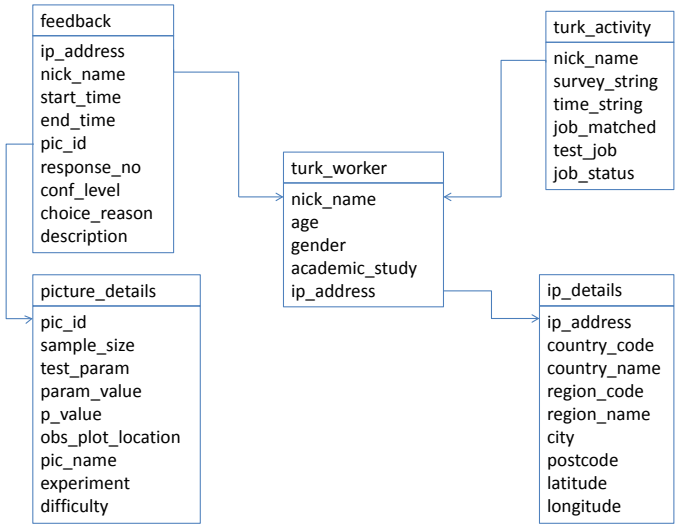


Figure 4.7 Relational database design for MTurk experiment data collection. The same database can be used for multiple turk experiments by keeping experiment information in picture_details table which contains information about the lineups.

The primary key in feedback table is produced by nick_name and pic_id since no workers are allowed to provide multiple feedbacks on the same lineup. In all other tables the first field names shown in Figure 4.7 are the primary keys. For the implementation of the web application we used MySQL database located on a local server securely accessible from public locations.

4.3.3 Data collection

The homepage of the web displays detailed explanation on how one can perform the task of evaluating the lineups. Several examples with possible answers are provided. It is possible to customize how the workers will proceed from the home page. There are two options; one is to allow them to continue the task where no trial is needed. The second option is to force them to try some lineup before joining the actual experiment. The trial feedbacks are not recorded. As per the requirement of Institutional Review Board (IRB) the workers need to provide the informed consent. For this they have to read and agree with specific IRB approved informed consent. The flow chart for this data collection sequence is shown in Figure 4.8.

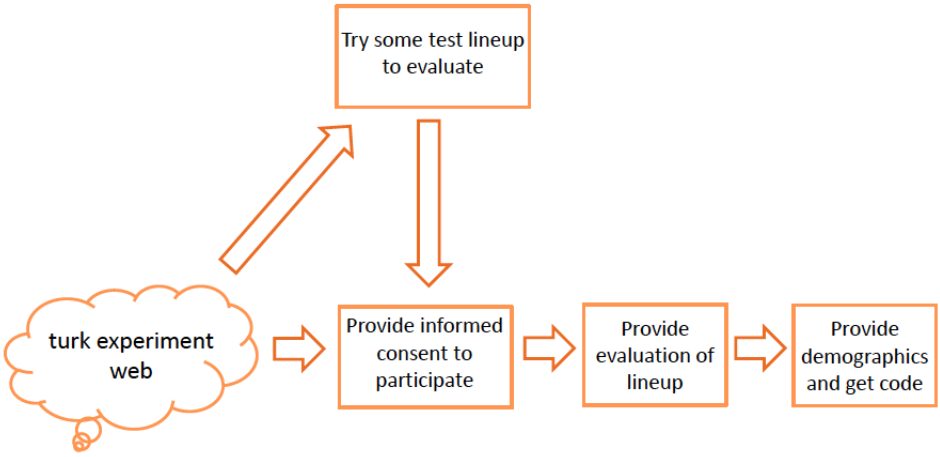


Figure 4.8 Data collection work flow shows that workers can try some test lineups before going to the live experiment after providing informed consent. This design gives the flexibility to make the trial mandatory, if needed, so that without having enough correct trial evaluations the actual participation can be prevented.

The default information that the system is designed to collect from each individual is shown in Table 4.1. Data received from each individual will be automatically saved in a secured mysql server.

Table 4.1 Default information the web application collects from each individual

Information	Description
Identification	Nick name or any ID to determine the responses of an Individual
Response number	The number of the plot on the lineup plot which the individual thinks the most different than other 19 plots.
Reason of of choice	Reason why the individual chooses the plot
Confidence level	Confidence level of individual choice
Age group	The age group where the individual belongs
Education	The highest level of education completed
Gender	Male or female
Geographic Location	This information is collected through the ip address of the individual computer
Time taken	Time taken for each response

4.3.4 Data Security and Validity

Since the experiment is based on online participations and related to monetary affairs, it is important to maintain certain security for the data collection. The attempt to provide random data or irrelevant information as well as harmful actions need to be prevented. Some cautious attempts are taken to add security to the data. Server side scripting language PHP is used to connect to database and access or save data. For transferring data from one form to another PHP session variables are used. Cookies are avoided carefully so that no important informations are saved in the cookies.

To prevent missing information and invalid input client side control is applied using JavaScripts. Most of the cases options or combo boxes are used instead of free text boxes to avoid invalid input. This also made the task easy to perform and convenient for the workers. For controlling some flow of the work such as showing various messages, JavaScripts are used as well. Careful consideration are made so that no important informations are stored in the java variable or function that could easily be revealed.

4.4 Managing Turk Task

Amazon (2010) Mechanical Turk website provides various features that are helpful to main-

tain and organize the task and manage the workers. Creating, posting, accepting or rejecting tasks and processing payment can be done from the MTurk system. It is also possible to screen out workers as needed. For example, none will be interested to recruit people who have a very bad reputation. This information can be obtained from each worker's previous work history of total number of accepted task out of total number of submitted task. The web application we developed is for presenting lineups in a flexible way the researcher may want as per the experimental design and collecting data. To get people to the web application we need to design a turk task in MTurk system.

4.4.1 Creating Task for Lineup Evaluation

To create a MTurk task some parameters have to be specified. The task descriptions need to be as simple as possible. This description makes the first impression about the task along with the amount of money to be paid for each task. Once the worker agrees to do the task they may come to the web site designed for lineup evaluation. Thus it is important to provide them with enough information so that they don't have to spend much time after coming to the web application. Some other parameters that need to be specified are discussed below.

- The number of workers to recruit has to be provided so that the turk task remains open until some specific number of people do the task. This is related to the sample size of the data we intend to collect on each lineup.

- Time allowed to finish each task has to be specified. Once a specific time period is set up the task will be expired after that time and the worker can't submit it anymore. The worker has to finish the task by that time period once he or she agrees to do the task.

- Qualifications needed to view and perform the task may be specified so that only desired participants can do the task. It depends on the population of interest as per the experimental design. For example one may only allow workers who have a history of doing at least 100 tasks and out of which a specific percent of tasks have been approved by other recruiters.

- Duration of the task is the period of time the task is available for workers to do. If the time period is over the task will not be available no matter whether all the required number of workers have participated or not. This duration is important as each payment has to be finalized by this time period. Otherwise, all the workers will be paid automatically after this time period no matter whether the the task is accepted or rejected. Some times it is helpful to set up a longer duration and see how long it takes before all the tasks get completed.

- Finally, the workers are redirected to the web site designed for lineup evaluation. MTurk workers should be informed clearly that they will be redirected to a new website from where they will provide feedback. The whole task including the instructions and procedures in the web application should be presented as per the task plan described in Section 4.2.5.

4.4.2 Accepting or Rejecting the Task

There is no hard and fast rule to follow while accepting or rejecting the tasks submitted by the MTurk workers. Since MTurk task are very simple and payment amount is very small the general convention is to pay every worker who submitted the task properly as instructed. The task should not be rejected based on the quality of work unless it clearly demonstrates that the task was not performed properly and irrelevant or garbage data were provided. It is very unusual to get garbage data from MTurk workers since they are well aware that this conduct may harm their reputation as a worker. So, proper caution needs to be taken while rejecting a task. Some times it may happen that the workers provided invalid data unintentionally. In that case we recommend to accept the task and pay for trying the task.

It is very difficult to detect whether a worker is giving data by properly inspecting the lineup or just at random. This can be monitored by putting a test lineup which is very easy to evaluate. The worker can be informed that there will be a test lineup which need to be correctly evaluated to make sure the task is accepted. This may prevent worker giving random data and it is easier to accept or reject tasks based on this criteria.

There may be mistakes made by the worker such as giving a wrong ID or a invalid input even though the whole task is properly done as instructed. In this case the task should be accepted and payment should be made. This encourages the workers and they appreciate the recruiters sincerity and provide good feedbacks. This does not happen frequently if qualified workers are recruited such as who had at least a certain number of percentage of tasks approved previously.

Other criteria to accept or reject the tasks may be more complex based on the experimental needs. For example one may accept the task if certain percentage of the easy lineups are correctly evaluated. This criteria can be used if multiple number of easy lineups are shown for evaluation. For example if there are 3 easy lineups and none of them were correctly evaluated it is most likely that the feedbacks were provided randomly. Some times it may happen that out of 10 lineups 3 difficult lineups were correctly evaluated but all the three easy lineups were wrongly evaluated. In that case the payment should be made since it is very unlikely that out of 10 lineups a total of 3 lineups would be correctly evaluated by just random response.

The payment of the workers is directly associated with the rejection of the work. If the task is not accepted the payment to the worker is denied and reasons for rejecting the tasks should be clear so that the worker doesn't get confused.

4.4.3 Managing Worker

Accepting the task and processing the payment in a timely manner may make a good impression about the task and the recruiter. This is helpful for the recruiter's reputation so that in future the recruiter can get the job done faster. Since the MTurk worker will always prefer the tasks posted by a good recruiter. Besides this we provide some good worker management strategy as below.

- Not all the workers put the same effort while doing the task. Some workers spend much time and perform the task sincerely and follow the instructions very carefully. That effort is observable in their responses and time taken for each evaluations as well as from the textual feedbacks such as writing choice reasons in details. We recommend paying bonus

to those workers in addition to the regular payment.

- If a work is not worth paying bonus payment we still recommend to appreciate each work. Even when the task gets rejected a proper explanation of rejection along with an appreciation for at least trying to participate the experiment is very useful for a recruiter's good image.

- Some workers may email to learn more about the experiment. Some times they may face technical troubles which needs to be taken care of promptly. A timely communication with worker is necessary and emails should be responded as promptly as possible. It is important as they can report any trouble to IRB which may create unpleasant issues with the human subject experiment.

- It is recommended to reject the task if it is not submitted as required. Some times this may produce dispute and the workers may become upset and still demand payment. We recommend to be strict to the decision and for any further administrative control there is an option to block the worker if necessary. After all this is a work place and it is expected that the worker maintain the work place environment.

4.5 Turk Experiment Data

We have performed 10 different experiments using the web application described in Section 4.3. Table 4.2 presents the tasks and payment details of all the 10 experiments. Experiment 1 was the first experiment and the largest number of tasks were rejected in that experiment. But this experiment provides valuable information about how to upgrade the web site to avoid invalid input and missing information. Thus in the later experiments we did not see a lot of tasks to get rejected.

The hourly payment rates shown in Table 4.2 are based on the time periods beginning from the times the tasks were accepted by the worker to the time they were submitted to the MTurk system. It is different from the time taken to actually evaluate the lineups. Since in between the workers need to go to the web application, get the codes after they finish the tasks and finally put the code to MTurk system. The pay rate is almost similar for all the experiment

Table 4.2 Amazon mechanical turk experiments and their properties. Duration in hours per 100 tasks show the popularity of some tasks compared to others.

	Experiment	Total Task		Average	Duration (hour)		Payment	Pay rate
Serial	description	submitted	rejected	time(min)	Actual	100 task	$/task	$/hour
1	Boxplot	406	106	10.68	146.48	36.08	0.50	2.81
2	Scatterplot	359	9	10.80	42.68	11.89	1.00	5.58
3	Contaminated plot	219	19	13.53	126.17	57.61	1.00	2.22
4	Polar vs Cartesian	110	10	20.65	11.65	10.59	1.00	2.91
5	Hist vs density	234	37	17.85	41.57	17.76	1.00	3.36
6	Violin vs boxplot	417	17	17.95	105.87	25.39	1.00	3.34
7	Group separation	106	6	16.13	5.15	4.86	1.00	3.72
8	Sine Illusion	101	1	16.52	78.38	77.60	1.00	3.63
9	Gene expression	103	3	12.47	11.27	10.94	0.50	2.41
10	Test normality	406	6	22.70	74.35	18.31	1.00	2.64

except for experiment 2. The payment amount was decided based on the MTurk standard of $6 per hour and we can see that for experiment 2. Based on the experiment 1 and 2, the payment for each task was fixed to $1, but in practice, it appears to be much less than that. It is because, workers spent long time in between lineups.

The duration of an experiment to finish depends on the time it was posted and the number of interested workers available on that time. It also depends on how attractive the experiment is in terms of appearance and payment. Even though the payment was similar for all the experiments, we observed some experiments were finished much faster than others. Experiment 8 took long time to finish which is about 77 hours for 100 tasks while experiment 7 took only 4.86 hours to finish 100 task.

Figure 4.9 shows the comparative durations of time for 100 tasks in each of the experiments. Even though experiment 8 (sine illusion) took longer to finish it got the least number of rejected tasks.

4.5.1 Data Cleaning

Data cleaning is different than accepting and rejecting the task. There may be some tasks for which payment is provided but yet the tasks may need to be excluded from the study. It is because there may be some participants who did not put in a best effort to identify the data plot in the lineup, but just randomly picked a plot to maximize their 'winnings'. We present following six suggestions about cleaning data from turk experiments where each worker

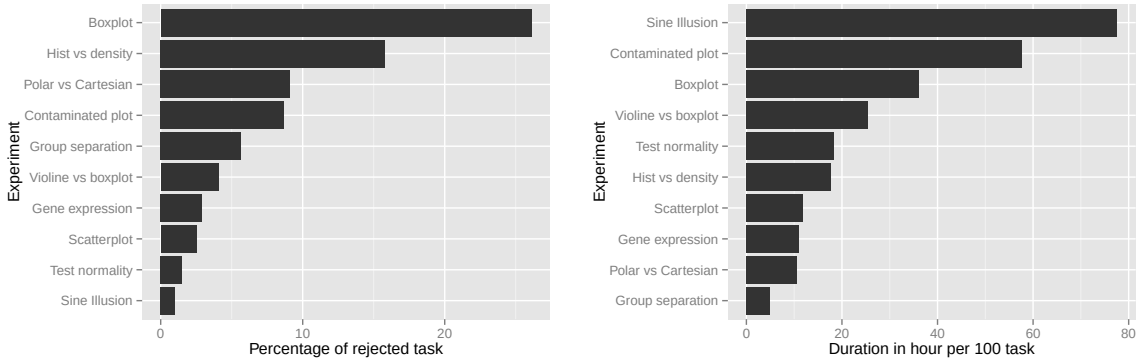


Figure 4.9 Percentage of rejected tasks and duration of each experiment in hour per 100 tasks for each of the 10 experiments. Most of the tasks got rejected for box plot experiment. Even though the sine illusion experiment took longest to finish the rejection rate is lowest for this experiment.

evaluates multiple lineups.

1. **include all participants** and their evaluations

2. exclude all participants' evaluations, who did **not** share their **demographic information** (age, gender education level – all three pieces of information are either missing or all present).

3. exclude participants' records, if **none of the evaluations** correctly identified the data plot – every participant was shown a range of 'easy' lineups.

4. include participants' records, if **at least 20 percent** of the evaluations are **correct** – based on ten evaluations per participants, two correct evaluations are significant evidence against a person just guessing

5. include participants' records, if at least **50% of all very easy lineups are correct**

6. use easy lineups as **reference charts**: sample one easy lineup from a person's records. If that lineup is evaluated **correctly**, include all (other) lineups of that person, otherwise exclude all lineup evaluations by this participant.

These screening criteria may be applied based on the specific design of the study and experimental evidence. For example Majumder et al. (2013b) used criteria 6 to clean the data.

4.5.2 Selection Bias

There is always a concern of selection biases with any online study. First of all in online study only those who use internet can participate. For a turk experiment the pool of participants even shrinks further to those who work in MTurk system.

Other biases may be due to the time the task is posted online, setting up specific qualification or even payment amount. A task which is posted at noon in central time in USA is less likely to be seen from Indian region since that is the midnight in that area. This may produce an artificial filter to the data which may come only from certain geographical location. If a specific qualification of the task is required not all the workers are able to view and perform that task. Also, if payment amount is very small many worker may not even review the task.

While selection bias may influence the results of some experimental study it is not much of an issue for experiment with lineups. In terms of demographic factors and geographic location MTurk provides much diversity in the participants as studied by Majumder et al. (2013a). Moreover it is observed that the human factors do not have any practical impact on the probability of correctly identifying the data plot in the lineup.

4.6 Conclusion

This paper presents a complete solution to the problem of designing a simulation experiment for lineup and recruiting people to get the lineups evaluated. The online application design provides features to control how the multiple lineups can be presented to an observer. Scalable Vector Graphics (SVG) are used for lineup so that observer can click on the lineup to pick certain plot. This also made the multiple plot selections from a lineup easier and convenient. The web site is used for multiple online experiments.

One of the main features of the web application is that it produces simple task for the worker but still retains many flexibilities to the researcher who need lineups to be evaluated as per complex experimental requirement. The design of this application allows recruiting people from any source not necessarily just from MTurk. The web application is now hosted on Iowa State University public domain (Majumder, 2013) and any number of experiments can be done

through this web site without changing the core of this application.

The next direction of this work is to make a complete package so that anyone can reproduce this web application, customize according to their needs and run the MTurk experiments as part of their research that involves lineups. This will also help lineup protocol to be used frequently in making inferential decisions since it will provide the required flexibilities and convenience.

We also intend to set up a web application for public use where researcher can put their lineup for online evaluation. The observer can be recruited from MTurk or any other sources including local lab participants.

Answer to the question in Figure 4.1 is 2.

CHAPTER 5. SUMMARY AND DISCUSSION

The three papers presented in this thesis established the validity of lineup protocol to use it as a tool for testing statistical hypothesis. Visual statistical inference is developed further by presenting the definitions of the terminologies. The methods of computing power of the visual test are proposed. Under some condition which is supported by experimental data, the power is obtained theoretically. A head to head comparison with the best available conventional test for regression slope is performed. The result suggests that visual test performs better when the effect size is large. For some super-visual individuals the performance is better even for small effect size. The influence of human factors on the visual test are examined and it is found that for some demographic and geographic factors the performance is better. But the practical impact of human factors is very negligible. Detailed procedures of human subject experiments are presented and the design of an web application to get lineups evaluated by human observer is provided. It offers various features that can be used by the researchers who intend to use lineup in decision making.

5.1 Future Work

This thesis opens up new areas of statistical research, new questions are now needed to be answered. The future research includes the in-depth analysis of these questions.

- *What are the characteristics for the best visual test statistics?* A follow up experiment was conducted using an eye-tracker to examine which patterns or features participants are cueing on in making their choices while evaluating a lineup Zhao et al. (2012). This gives important hints about the effective visual test statistics. Our future research involves characterizing the features of a visual test statistic in terms of color, orientation and plot

type that produces best power.

- *Can inference be included in basic statistics curriculum?* Our experimental data suggest that visual inference is a very intuitive decision making process which does not reacquire advanced knowledge on mathematics or statistics. The statistical inferential technique could be included in elementary school curriculum to prepare them well ahead for advanced statistical inferential technique. This requires more experiments and analysis. But it could a broad area of research for statistics education.

- *What is the best lineup size?* The experimental setups in Chapter 2 are based on the lineup of size 20. A theoretical justification of picking lineup size 20 is also given. The lineup size is associated with the type-I error of the test. But how actually a different size affect the performance of the observer is yet to be examined by proper experiment. A lineup of size 10 may be faster to evaluate compared to a lineup of size 30. More research is needed to learn how the size of the lineup helps or affects the evaluation process.

- *Can visual inference be used in big data problem?* This dissertation work shows promises for lineup protocol to be used in big data scenarios where it is hard to find a conventional way of making inferential decisions. Some of the applications of the lineups are shown in Hofmann et al. (2012) and Roy Chowdhury et al. (2012). It would be interesting to investigate the application further with a situation where only visual inference would be a solution for inference. The future research may explore this with big data.

- *Does a person well trained on statistical graphics yield better power?*We have observed subject to subject variability in the performance of evaluating a lineup. If an observer is well trained on statistical graphics, it may help an observer to critically examine the lineup in the direction of the question asked while showing the lineup. The skilled observers may perform better than the unskilled observers who do not have any exposure to or knowledge about statistical graphics. But this difference in performance of skilled and unskilled observer needs to be examined using controlled experimental set up.

- *Does multiple response option yield better power?* While in most of our experiments only

one response was taken from each observer, in some of the experiments observers were allowed to pick multiple plots from a lineup. Apparently, flexibility of multiple responses did not show any improvement in observer performance. This needs to be explored further to study if the multiple responses indeed improve the power of the visual inference.

5.1.1 Mathematical Framework of Visual Inference

In visual inferential procedure, the test statistic is a plot which is not a random variable in classical definition. But it is a function of data and hence inherits uncertainty due to sampling variations. Thus it is required to broaden the scope of the definition of random variable to bring the statistical inference world under a more general mathematical framework.

Mathematicians have done a lot of work to give the mathematical framework for statistical methods and procedures. But statistical techniques are expanding its horizon and due to practical necessity it is not limiting its fundamental base on some limited mathematical framework. Visual statistical inference is such an example. This is where we need to focus on how we can come up with a more general mathematical framework for statistical inference as a whole.

As we see in chapter 2, the concept of the power of a visual test is no more depending on some limited regularity conditions. Rather, the fundamental classical assumption for uniformly most powerful (UMP) test made it more vulnerable in front of newly developed visual inference procedure. In this perspective we need to focus on making the current mathematical framework more general to incorporate broader area of inferential statistics. I intend to work on this area in near future.

5.2 Final Remark

Visual statistical Inference offers promises when there is no formal way of testing hypothesis available. Even when a conventional test exists, this modern age of data driven society produces such an abundance of data that it is very easy to obtain statistical significance. Because with huge amount of data, one may reduce the random error variability to arbitrarily small. This brings new challenges with conventional tests since we only care for practical significance. Visual statistical inference may provide practical significance.

APPENDIX A. SUPPLEMENTARY MATERIALS OF CHAPTER 2

The materials in this document supplement the information presented in the manuscript "Validation of Visual Statistical Inference, Applied to Linear Models". Section A.1 presents the proof of the Lemma 2.3.1 in the manuscript. Section A.2 describes how the lineups are presented to the subjects for evaluation. The data cleaning process is described in Section A.3, supplementing Section 2.6.1 of the manuscript. A detailed discussion on how much the sample of null plots might affect the observer's choice is in Section A.4, supplementing a summary given in Section 2.6.7 of the manuscript. Section A.5 contains more discussion about Type-III error, and supplements Section 2.6.8 of the manuscript.

A.1 Proof of the Lemma

The proof of the Lemma 2.3.1 in the manuscript is shown below;

Proof. By definition

$$p_D = Pr\left(|t| \geq t_{obs} \mid H_0\right) = 1 - F_{|t|}(t_{obs}) \quad \Rightarrow \quad |t_{obs}| = F_{|t|}^{-1}(1 - p_D)$$

Then the distribution function of the p-value, p_D, under H_0, is uniform, since:

$$\begin{aligned} F_{p_D}(p) &= Pr(p_D \leq p) = 1 - Pr(1 - p_D \leq 1 - p) \\ &= 1 - Pr\left(F_{|t|}^{-1}(1 - p_D) \leq F_{|t|}^{-1}(1 - p)\right) \\ &= 1 - Pr\left(|t_{obs}| \leq F_{|t|}^{-1}(1 - p)\right) \\ &= 1 - F_{|t|}\left(F_{|t|}^{-1}(1 - p)\right) = p \ ; \text{ under } H_0 \end{aligned} \tag{A.1}$$

Let $p_{0,i}$, $i = 1, ..., m-1$ denote the p-values associated with data corresponding to the $m-1$ null plots. Since this data is generated consistently with the null hypothesis, the p-values are

independent and follow a standard Uniform distribution, $p_{i,0} \sim U[0,1], i = 1, ..., m-1$. The minimum $p_0 = \min_{1 \leq i \leq m-1} \ p_{0,i}$ then follows a Beta distribution with shape parameters 1 and $m-1$, and corresponding distribution function

$$F_{p_0}(x) = 1 - (1-x)^{m-1} \text{ for } x \in [0,1].$$

Thus

$$\begin{aligned} P(p_D < p_0) &= 1 - P(p_0 \leq p_D) = 1 - \int_0^1 P(p_0 \leq p_D \mid p_D = t) f_{p_D}(t) dt \\ &= 1 - \int_0^1 F_{p_0}(t) f_{p_D}(t) dt = 1 - \int_0^1 f_{p_D}(t) dt + \int_0^1 (1-t)^{m-1} f_{p_D}(t) dt \\ &= E\left[(1-p_D)^{m-1}\right]. \end{aligned}$$

□

A.2 Selection of Lineups for each subject

Table A.1 shows the selection process of the lineups for the subjects, across the experimental design parameters of experiment 1 (Section 2.5.1 of the manuscript), as required to obtain a margin of error of 0.05. The lineups are divided into four groups – easy, medium, hard and mixed – based on the parameter combinations shown in the table. The number of evaluations, along with the number of lineups, and the number of lineups from each category that a single subject would get, are shown in the table. Note that, every subject saw a block of 10 lineups, selected across these groups, including at least 1 easy lineup, and possibly 2 if one was drawn from the mixed group. These ideal sample sizes were generated using a goal of obtaining a margin of error no bigger than 0.05. For example, a lineup with sample size = 100, standard error = 5 and slope parameter = 3 requires 203 evaluations so that the proportion correct can be estimated with margin of error of 0.05 following the procedures described in the manuscript. A total of 300 subjects would provide a total of 3000 evaluations with this plan. Table A.2 shows the number of subjects actually participating in experiment 1 is 424 which is much higher than 300, but the number after cleaning was 239.

Table A.1 Ideal numbers for different experimental design parameters for exaperiment 1 (Section 2.5.1 of manuscript) in order to obtain a margin of error of 0.05. These numbers are used to choose a sample of 10 lineups for each subject.

Difficulty level	parameter combination			Number of evaluations required (n_γ)	Total number of lineups	Number of lineups randomly shown
	n	σ	β			
easy	100	5	8	1	12	1
	100	12	16	1		
	300	5	5	1		
	300	12	10	1		
medium	100	5	3	203	9	2
	300	5	2, 3	97, 1		
hard	100	12	3, 8, 10	277, 126, 23	18	6
	300	5	1	371		
	300	12	3, 5	375, 74		
mixed	100	5	1, 5, 0	214, 2, 73	21	1
	100	12	1	100		
	300	5	0	73		
	300	12	7, 1	2, 152		
Total					60	10

A.3 Data Cleaning

Amazon Turk is a relatively new source of subjects for experiments. The workers ("turkers") are paid, minimal amounts for their efforts, on par with conventional human subject experiments. Most turkers make an effort to complete tasks as requested, but some turkers do not take the task seriously. Our procedure for ensuring that reliable data was available for analysis was to provide one very easy lineup, one in which the observed data plot stands out as being very different from the null plots. The subject was informed that an easy lineup would be used to accept their evaluations. If that lineup is evaluated **correctly**, we include all (other) lineups of that subject, otherwise we exclude all lineup evaluations by this participant. Table A.2 displays number of subjects and their total evaluations before and after cleaning the data.

After cleaning the data, for experiment 1, we did not have 300 subjects that we planned for. It was decided that more data was not needed, though, because the estimated margin of error with the 239 subjects was close to 0.05. This can be seen from the bootstrap confidence band in Figure 2.6 of the manuscript. With experiments 2 and 3 there was sufficient data even after cleaning. Part of the success with the later experiments comes from the researchers developing a reputation on Amazon for providing a good task and reliable payment, which means the

Table A.2 Number of unique subjects and their total feedbacks before and after data cleaning. Note that the number of male and female participants may not add up to the number of subjects, due to some participants declining to provide demographic information.

Data	Experiment 1				Experiment 2				Experiment 3			
cleaning	Subj	Male	Fem	Total	Subj	Male	Fem	Total	Subj	Male	Fem	Total
before	424	226	180	4516	386	199	182	4330	242	158	79	2565
after	239	121	107	2249	351	185	164	3636	155	103	52	1511

reliable turkers look specifically for these tasks.

A.4 How much do null plots affect the choice?

It is discussed in the Section 2.6.7 of the manuscript that p-values can be used to quantify the similarity of the visual pattern in the plots used for the simulation experiments. Based on this, we explore more details on how much the null plots affect the choice made by the subjects.

We have seen that the subjects tend to pick the plot in the lineup that has the lowest p-value (Figure 2.10 in the mauscript). What we are also interested in is how this pick is affected by the distribution of p-values of other plots in the lineup, particularly the p-value of the null plot with the strongest structure. If there is a null plot with a small p-value, or one close to that of the actual data plot, we would expect that subjects have a harder time detecting the actual data plot. Figure A.1 investigates this. The difference between the p-value of the actual data is compared with the lowest from the null plots. This is plotted horizontally, and the proportion correct is plotted vertically. Negative values indicate lineups where the actual data plot had a smaller p-value than the minimum of the null plots. In experiment 1 (boxplots) there were a lot of lineups where the actual data plot had the smallest p-value, but only just. This caused quite some confusion for subjects, as seen because the variability in the proportion correct is huge for these lineups. Similarly large variability in correctness can be seen in the results of experiment 2 (scatterplots) except that the greater range of of differences in p-values shows the strength of subject's ability to pick the plot with most structure. Figure 2.10 in the manuscript shed some more light on this story: when there is a big difference between the p-values (eg experiment 1, $\beta > 7$) the subjects as one force chose the same plot. When there is less difference the

distribution of counts is much more evenly spread between plots (eg experiment 1, $\beta = 1$).

In practice, the p-value is not going to be a valid way to compare plots. Rather metrics that can measure how graphical elements from one plot to another are perceived similarly are needed. This is investigated in Roy Chowdhury (2012). Here, numerical measures of the similarity between plots are proposed to provide quality metrics for lineups.

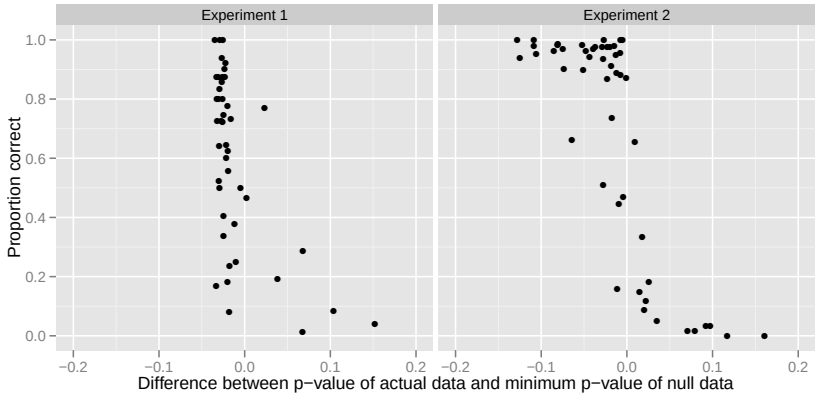


Figure A.1 Scatter plot of difference between the data plot's p-value and the smallest p-value of the null plots vs proportion correct. Negative differences indicate the p-value of the actual data plot are smaller than those of all of the null plots. Difference close to zero shows a wide range in the proportion correct, suggesting that when at least one null plot has structure almost as strong as the actual data plot, subjects had a difficult time in making their choice.

A.5 Type III error

Figure 2.6 in the manuscript indicates that Type III error might be occurring in experiment 3: correct identification of the actual data plot is not positively associated with effect size. Teasing this out of the results is possible by looking at the reasons participants gave for their choices. Participants were provided with four possible reasons to use for their choice:

1. Most different plot
2. Visible trend
3. Clustering visible
4. Other

with the possibility to use more than one. The task requested subjects to identify the plot that had the largest slope, which would correspond to choosing "visible trend" (2) as the reason for their choice. Reasons 1 or 3 would be indicative of Type III error. Figure A.2 explores the reasons subjects gave for their choices. If there were no Type III errors committed, we would expect that people overwhelmingly using "visible trend" as their reason, or at least, when they use this reason they overwhelmingly correctly choose the actual data plot. This is not what we see. At left, are the reasons subjects gave for their choices — 123 means that they gave all three reasons. The horizontal axis shows proportion of times that subjects correctly chose the actual data plot, and the reasons are sorted from most accurate to least accurate. The size of the point corresponds to the number of subjects putting this as the reason. Subjects that chose all three reasons almost always chose the actual data plot. This was followed by using 1 and 3, and then 1 and 4. The most common reasons given were reasons 1-3 individually, and the accuracy for these reasons ranged from 75% for reason "most different plot" to 60% for "visible trend". At right is a simplified view, containing just the four possible reasons – if the subject chose one of these, regardless if they also chose another reason it is counted. "Visible trend" comes in third. This is strong evidence that for many subjects even though they are correctly choosing the data plot, often they are cueing to other structure in the plot than the trend, making a Type III error.

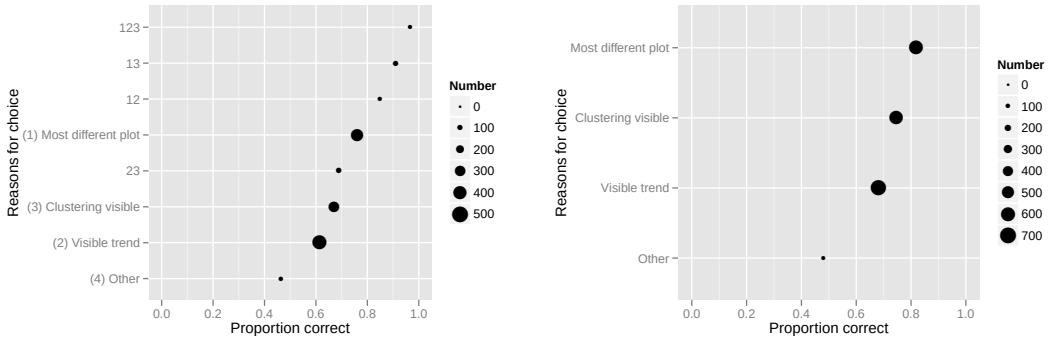


Figure A.2 Reasons of plot choices vs proportion of times the subjects correctly chose the actual data plot for experiment 3 that examines the occurrence of Type III error. At left, all subjects' choices are shown, and reason 123 means all three reasons are used. At right, if the subject used a reason, regardless if they also used more than this reason, they are counted. Size of the point corresponds to the number of subjects using that reason.

APPENDIX B. SUPPLEMENTARY MATERIALS OF CHAPTER 3

B.1 How People Pick the Data Plot

In each of the experiments observers were asked to choose the reason for their selection of a particular data plot. Majumder et al. (2013b) showed that these choice reasons may provide the clue on how people picks the data plot in the lineup. To investigate this further we added free text input option in experiment 9 for their choice reason instead of some fixed reasons to select from. This allowed observers to write whatever they think their reasons for choice are. Figure B.1 shows the words used to explain their reasons for choice. The most common words used to explain their choice are points and green which indicates the use of two important features of the grammar of graphics (Wilkinson, 1999; Wickham, 2009). One is the indicator of geometric shape and the other is the aesthetics. Spread, steepest, line and apart are some other important words used frequently. Spread and apart are indicative of variability in the data. Steepest and line indicates some sort of systematic pattern in the data.

Figure B.1 Words used to explain the reasons for selection of data plot in a lineup show what features of a lineup may help a non-statistician to evaluate it. Larger font indicate more people choosing that word. Different color is used just to separate the words.

Figure B.1 also shows some insight about peoples way of reading a plot. We notice that variability in the data, geometric shapes and aesthetics used to generate plots and existence of any systematic pattern in the plot are some of the important features revealed from the figure. These features are commonly used by human brain to examine and compare plots. Notice that these features may be specific to this particular experiment. There can be different other features people may use to evaluate a lineup depending on the situation. But with these words we get a general idea on how people may think while evaluating a lineup.

Turk workers are not necessarily trained on statistics or aware of specific terms used in statistics or statistical plots let alone having knowledge about grammar of graphics. It is interesting that people explain things that have specific definitions and meaning in the literature. For example, some keywords like spread, apart should be analogous to larger variability while together, close may be for indicating smaller variability. Thus the perception from a statistical graphics is intuitive and human intelligence learn this even without having specific training. This is why visual inference can be used as a tool for teaching inference in the basic statistics classes.

B.2 Some Plots of Exploratory Data Analysis

This section includes some plots showing results from exploratory data analysis. The results are discussed in Section 3.4.

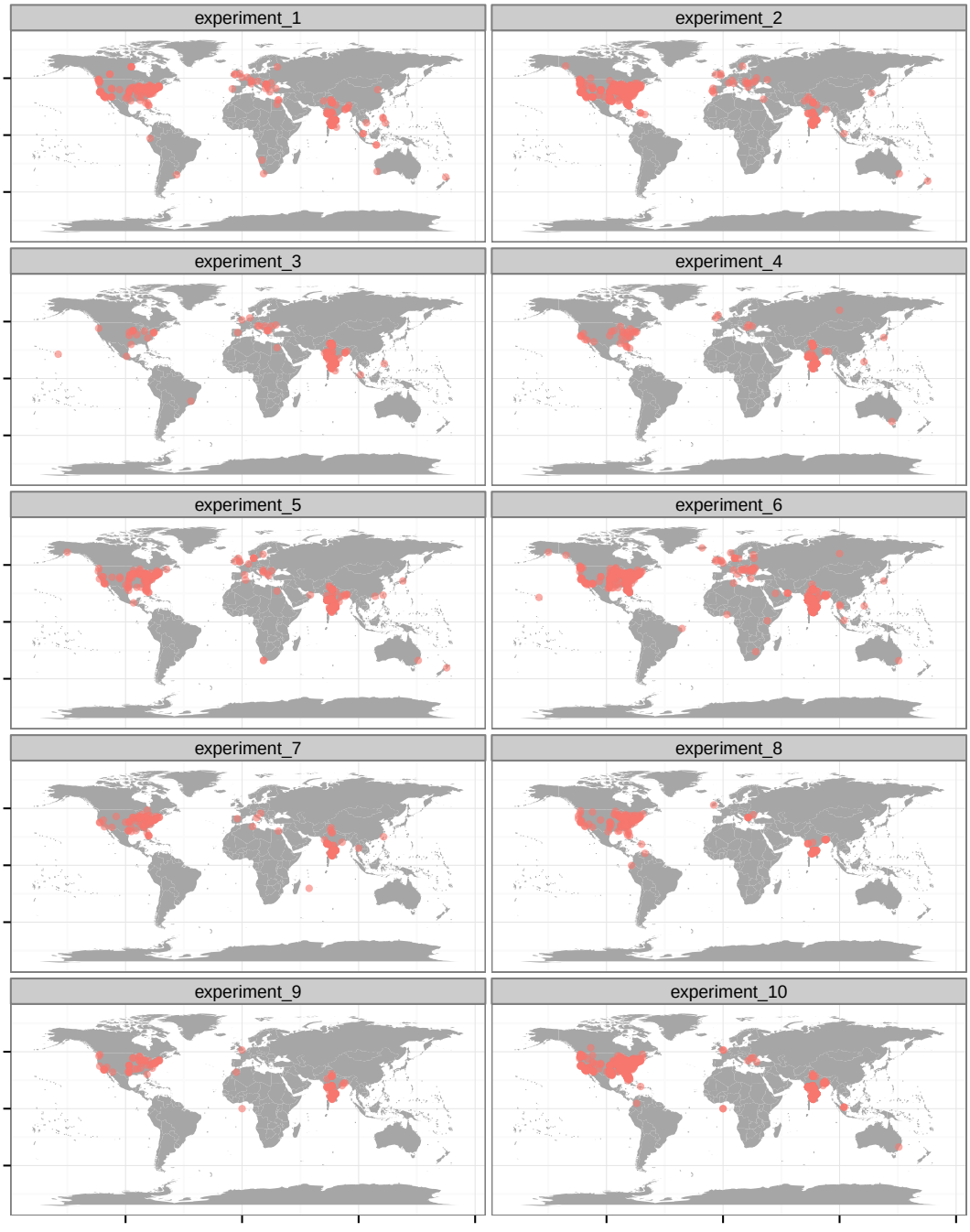


Figure B.2 World maps showing where the participants are coming from for all the 10 experiments.

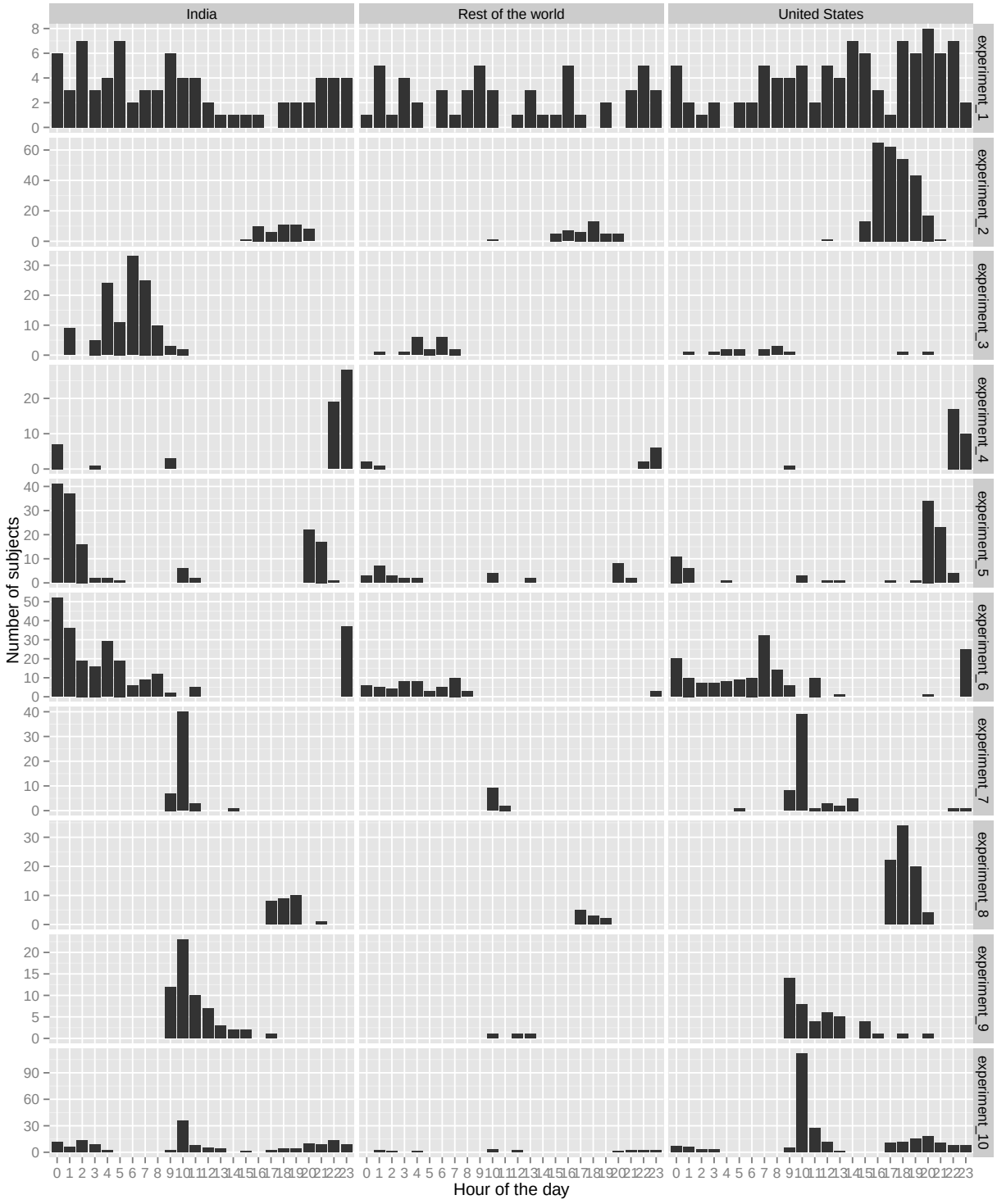


Figure B.3 Number of participants by time of the day feedbacks received (central time). Experiment 1 shows MTurk workers participated the experiments around the clock. Other experiments did not take a whole day to finish. For experiment 3 most of the participants are from India because of timing. No matter when the experiment is started, subjects from India shows participations. For United States, subjects participated if the experiment is not in the mid night, except for experiment 6.

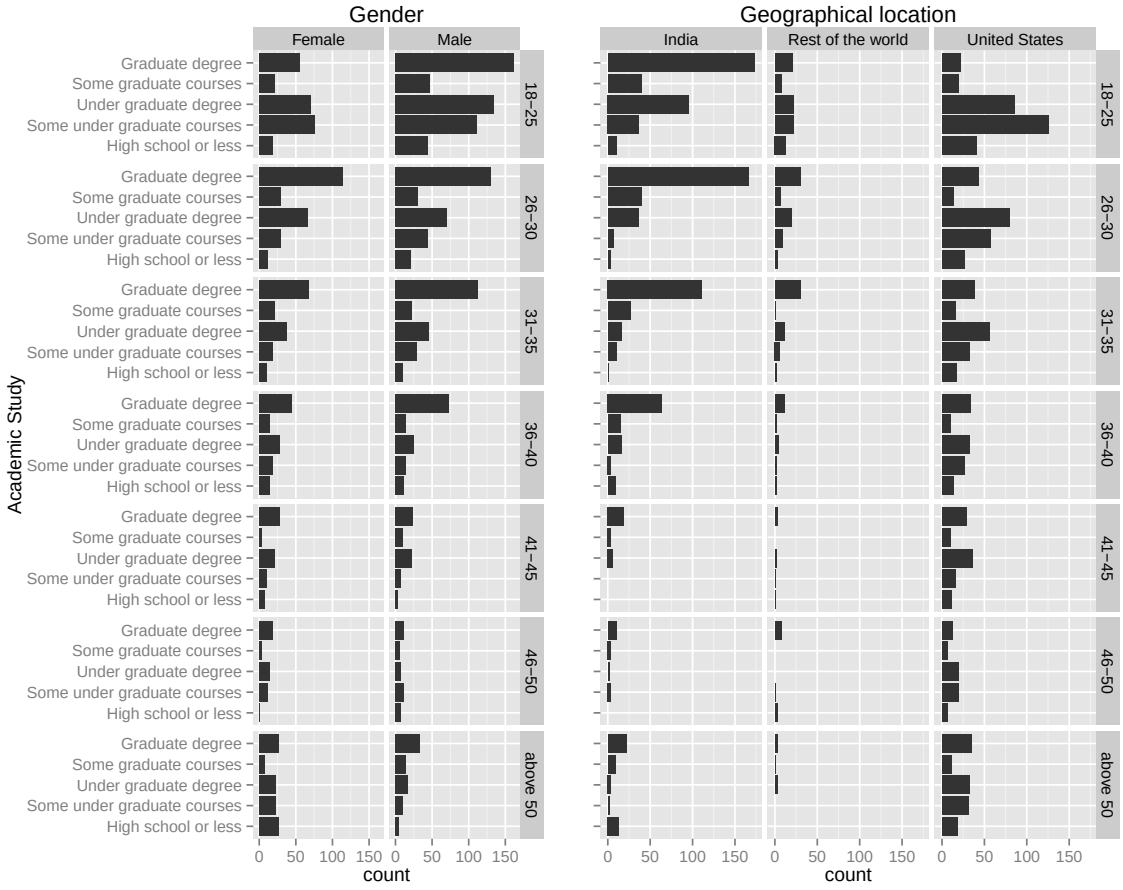


Figure B.4 Countrywise distribution of age and academic levels of the MTurk workers participating the experiments shows the diversity of the subjects in all the demographic aspect. Almost equal number of male and female subjects participated the online experiments.

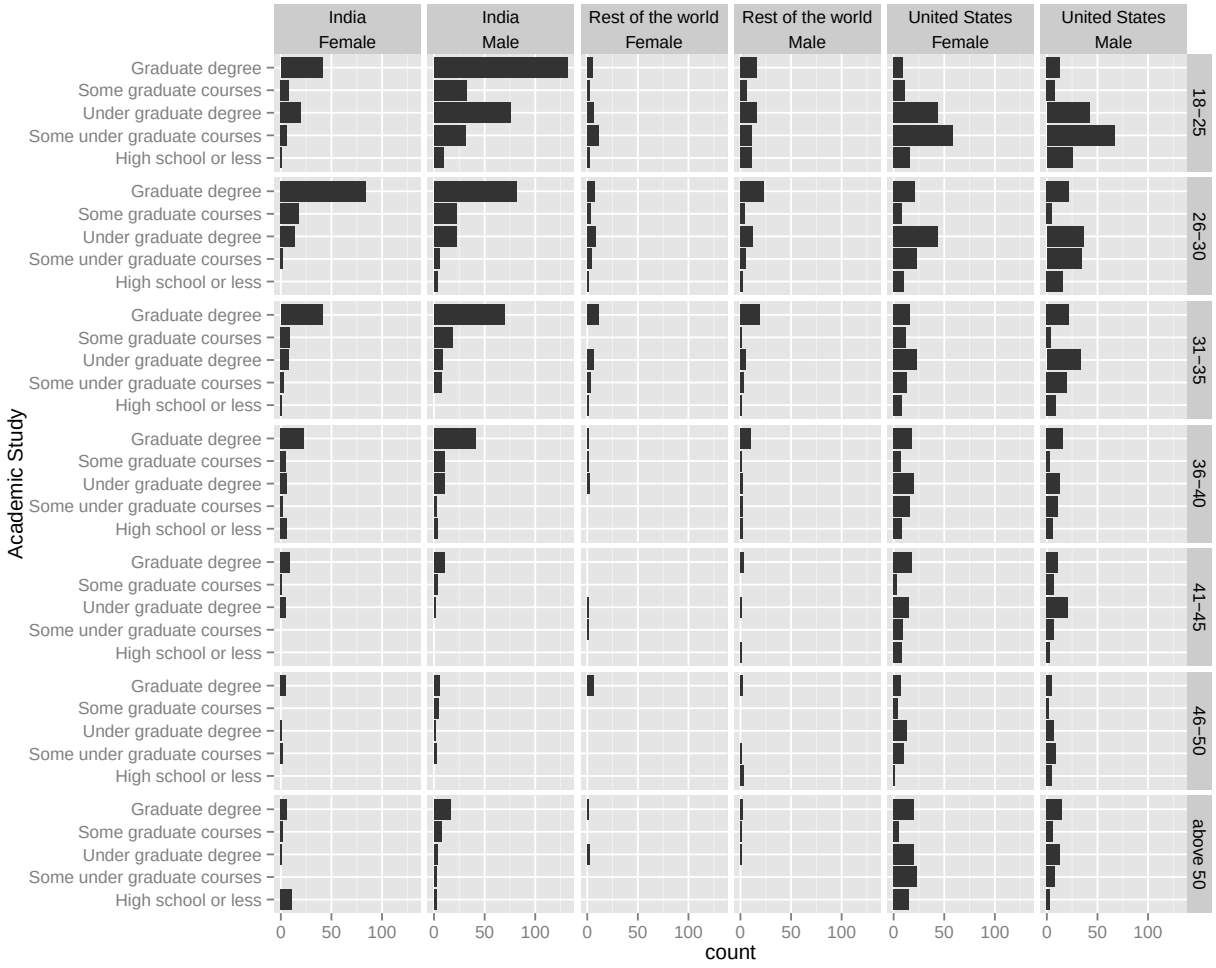


Figure B.5 Countrywise distribution of age and academic levels of the MTurk workers participating the experiments shows the diversity of the subjects in all the demographic aspect. Male and female participants differ in India specially for agelevel 18-25. For United States number of participants are similar beyond age 40 while few number of participants coming from India beyond that age.

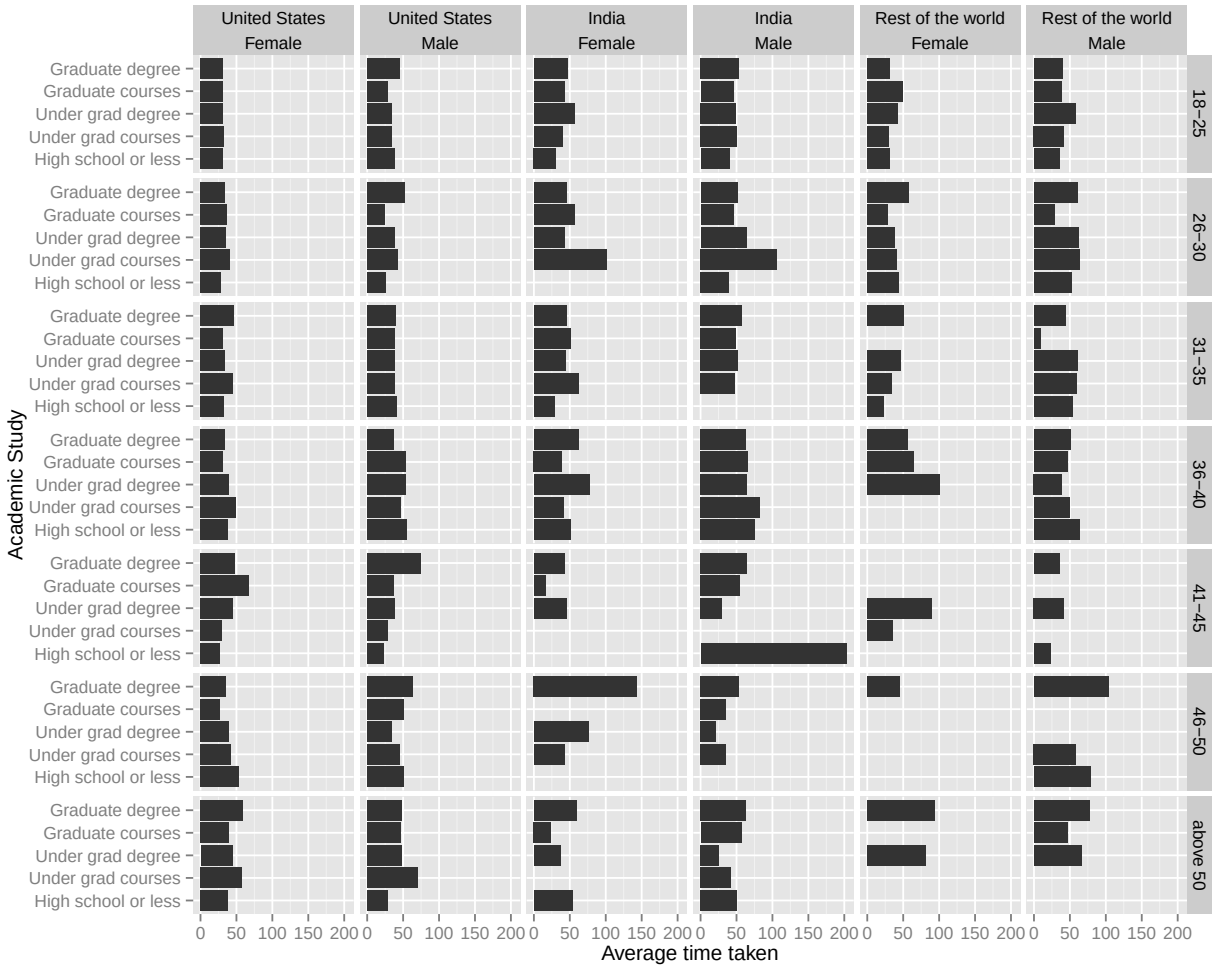


Figure B.6 Countrywise average time taken for different age and academic levels of the MTurk workers participating the experiments shows that the demographic factors may not have effect on time taken.

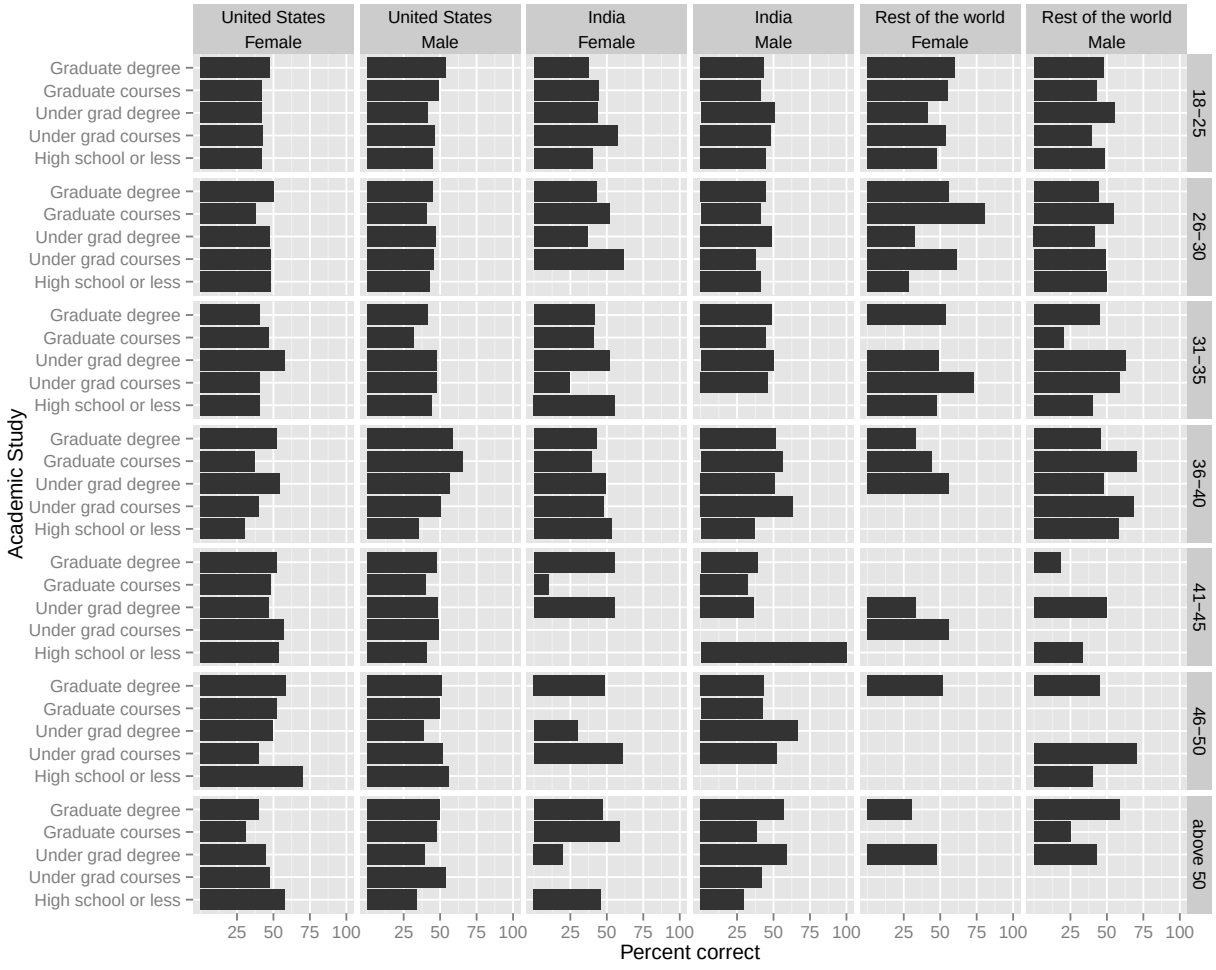


Figure B.7 Countrywise percentage of correct responses for different age and academic levels of the MTurk workers participating the experiments shows that the demographic factors may not have effect on the percentage of correct responses.

B.3 Electoral Building Lineups and Results

Five lineups were shown to (different) Amazon Turk workers in an experiment. They all were created as described in the introduction of the paper. In order to not bias observers, no context information was given about how these plots were created or what data they were displaying. This also made it necessary to slightly more stylize the display.

lineup	#1	#2	#3	#4	#5
# correct/ #evaluation	12/72	11/66	5/74	14/72	19/57
p-value	0.00023	0.00041	0.31	1.2e-05	1.9e-11
data panel	$3 \cdot 4 + 1$	$2^4 + 1$	$4^2 + 2$	$12 + \sqrt{25}$	$2^3 - 7$

Table B.1 Overview of all choices by observers for each of the lineups. The correct choice is bolded. In most lineups there are null plots that were picked more often by observers, but the actual result is among the plots being picked most often, indicating that there is some indication that the election result is not completely consistent with the polls.

	panel chosen																			
Lineup	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
# 1	2	2	0	10	2	2	6	23	1	1	0	1	**12**	3	3	1	0	1	1	1
# 2	0	16	1	1	5	1	0	8	0	2	0	0	0	4	2	1	**11**	1	0	13
# 3	7	26	0	2	0	5	3	0	2	1	0	4	0	0	2	0	9	**5**	0	6
# 4	0	0	0	2	0	0	0	3	1	10	2	18	1	0	4	2	**14**	0	13	0
# 5	**19**	1	4	1	0	1	0	12	0	0	0	4	1	0	0	12	1	1	0	0

BIBLIOGRAPHY

Amazon (2010). Mechanical Turk. https://www.mturk.com/mturk/welcome.

Atwood, S. E., O'Rourke, J. A., Peiffer, G. A., Yin, T., Majumder, M., Zhang, C., Cianzio, S., Hill, J. H., Cook, D., Whitham, S. A., Shoemaker, R. C., and Graham, M. A. (2013). Gmrpa3 and the iron efficiency stress response in soybean. *Plant Cell and Environment*, 3(1):3. in press.

Bates, D., Maechler, M., and Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-42.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F., and Wickham, H. (2009). Statistical inference for exploratory data analysis and model diagnostics. *Royal Society Philosophical Transactions A*, 367(1906):4361–4383.

Buja, A. and Rolke, W. (2011). Calibration for simultaneity: (re)sampling methods for simultaneous inference with applications to function estimation and functional data. Unpublished manuscript.

Butler, K. and Stephens, M. (1993). The distribution of a sum of binomial random variables. *Tech Rep 467 Stanford University Stanford Calif USA April 1993 prepared for the Office of Naval Research*, 0(467).

Cleveland, W. S. and McGill, R. (1984). Graphical perception: Theory, experimentation and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554.

Cook, R. D. and Weisberg, S. (1999). *Applied Regression Including Computing and Graphics*. Wiley Series in Probability and Statistics.

Heer, J. and Bostock, M. (2010). Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI Í10, pages 203–212, New York, NY, USA. ACM.

Hofmann, H., Follett, L., Majumder, M., and Cook, D. (2012). Graphical tests for power comparison of competing designs. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2441–2448.

Majumder, M. (2013). A web application for turk experiment. http://www.public.iastate.edu/~mahbub/feedback_turk11/homepage.html.

Majumder, M., Hofmann, H., and Cook, D. (2013a). Human factors influencing visual statistical inference. *Sociological Methodology*. To be submitted.

Majumder, M., Hofmann, H., and Cook, D. (2013b). Validation of visual statistical inference, applied to linear models. *Journal of the American Statistical Association*. Accepted for publication.

Mason, W. and Suri, S. (2012). Conducting behavioral research on amazon's mechanical turk. *Behavior Research Methods*, 44(1):1–23.

Mosteller, F. (1948). A k-Sample Slippage Test for an Extreme Population. *The Annals of Mathematical Statistics*, 19(1):58–65.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

R Development Core Team (2012). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Roy Chowdhury, N., Cook, D., Hofmann, H., and Majumder, M. (2011). Visual statistical inference for large p, small n data. In *JSM Proceedings*, pages 4436–4446, Alexandria, VA. Section on Statistical Graphics, American Statistical Association.

Roy Chowdhury, N., Cook, D., Hofmann, H., Majumder, M., and Zhao, Y. (2012). Where's waldo: Looking closely at a lineup. Technical Report 2, Iowa State University, Department of Statistics.

Simkin, D. and Hastie, R. (1987). An information processing analysis of graph perception. *Journal of the American Statistical Association*, 82:454–465.

Spence, I. and Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, 6:61–77.

Suri, S. and Watts, D. J. (2010). Cooperation and contagion in networked public goods experiments. *CoRR*, abs/1008.1276.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. useR. Springer.

Wilkinson, L. (1999). *The Grammar of Graphics*. NY: Springer, New York.

Yin, T., Majumder, M., Roy Chowdhury, N., Cook, D., Shoemaker, R., and Graham, M. (2013). Visual mining and inference for rna-seq data. *Journal of Data Mining in Genomics & Proteomics*. submitted.

Zhao, Y., Cook, D., Hofmann, H., Majumder, M., and Roy Chowdhury, N. (2012). Mind reading using an eyetracker to see how people are looking at lineups. Technical Report 10, Iowa State University, Department of Statistics.